The copyright © of this thesis belongs to its rightful author and/or other copyright owner. Copies can be accessed and downloaded for non-commercial or learning purposes without any charge and permission. The thesis cannot be reproduced or quoted as a whole without the permission from its rightful owner. No alteration or changes in format is allowed without permission from its rightful owner.

AN ENHANCED SEQUENTIAL EXCEPTION TECHNIQUE FOR SEMANTIC-BASED TEXT ANOMALY DETECTION

MOHAMMED AHMED TAIYE

Universiti Utara Malaysia

DOCTOR OF PHILOSOPHY
UNIVERSITI UTARA MALAYSIA
2019

AN ENHANCED SEQUENTIAL EXCEPTION TECHNIQUE FOR SEMANTIC-BASED TEXT ANOMALY DETECTION

By

MOHAMMED AHMED TAIYE

Universiti Utara Malaysia

Thesis Submitted to

Awang Had Salleh Graduate School of Arts and Sciences,

Universiti Utara Malaysia

In Fulfilment of the Requirement for the Degree of Doctor of Philosophy

Awang Had Salleh
Graduate School
of Arts And Sciences

Universiti Utara Malaysia

PERAKUAN KERJA TESIS / DISERTASI
(Certification of thesis / dissertation)

Kami, yang bertandatangan, memperakukan bahawa
(We, the undersigned, certify that)

MOHAMMED AHMED TAIYE

calon untuk Ijazah
(candidate for the degree of) **PhD**

telah mengemukakan tesis / disertasi yang bertajuk:
(has presented his/her thesis / dissertation of the following title):

"AN ENHANCED SEQUENTIAL EXCEPTION TECHNIQUE FOR SEMANTIC-BASED TEXT ANOMALY DETECTION"

seperti yang tercatat di muka surat tajuk dan kulit tesis / disertasi.
(as it appears on the title page and front cover of the thesis / dissertation).

Bahawa tesis/disertasi tersebut boleh diterima dari segi bentuk serta kandungan dan meliputi bidang ilmu dengan memuaskan, sebagaimana yang ditunjukkan oleh calon dalam ujian lisan yang diadakan pada : ***02 Mei 2019.***
That the said thesis/dissertation is acceptable in form and content and displays a satisfactory knowledge of the field of study as demonstrated by the candidate through an oral examination held on: ***May 02, 2019.***

Pengerusi Viva: *(Chairman for VIVA)*	**Prof. Dr. Huda Hj Ibrahim**	Tandatangan *(Signature)*
Pemeriksa Luar: *(External Examiner)*	**Assoc. Prof. Dr. Shuzlina Abdul Rahman**	Tandatangan *(Signature)*
Pemeriksa Dalam: *(Internal Examiner)*	**Dr. Mohamad Farhan Mohamad Mohsin**	Tandatangan *(Signature)*
Nama Penyelia/Penyelia-penyelia: *(Name of Supervisor/Supervisors)*	**Assoc. Prof. Dr. Siti Sakira Kamaruddin**	Tandatangan *(Signature)*
Nama Penyelia/Penyelia-penyelia: *(Name of Supervisor/Supervisors)*	**Dr. Farzana Kabir Ahmad**	Tandatangan *(Signature)*

Tarikh:
(Date) ***May 02, 2019***

Permission to Use

In presenting this study in fulfilment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the Universiti Library may make it freely available for inspection. I further agree that permission for the copying of this study in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor or, in her absent, by the Dean of Awang Had Salleh Graduate School of Arts and Sciences. It is understood that any copying or publication or use of this study or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to University Utara Malaysia for any scholarly use which may be made of any material from my study.

Requests for permission to copy or to make other use of materials in this study, in whole or in part should be addressed to:

Universiti Utara Malaysia

Dean of Awang Had Salleh Graduate School of Arts and Sciences
UUM College of Arts and Sciences
Universiti Utara Malaysia
06010 UUM Sintok

Acknowledgments

All praise to Almighty Allah (SWT) who gave me courage and patience to carry out this work. Alhamdulillah.

My sincere appreciation and gratitude goes to my supervisors, Professor Madya Dr. Siti Sakira Kamurddin and Dr. Farzana Kabir Ahmed. My appreciation also goes to my scholarship guarantor and lecturer Dr. Norliza Binti Katuk and My Late supervisor Professor Mohammed Syazwan B. Abdullah for their academic guidance, support and encouragement. Not forgotten the appointed examiners who have given valuable comments to improve my study.

I wish to express my gratitude to the academic and supporting staff of School of Computing, Universiti Utara Malaysia for all the assistance rendered during my studies. I wish to thank all my friends, whose continuous discussions and, support greatly helped in this research.

I would like to dedicate this work to my twin brother Mohammed Kehinde Mohammed and my son Mohammed Nabeel, thank you. My wife Maryam Olaoti Shehu Mohammed for the patience and support. my caring elder siblings Ibrahim Mohammed and my Lovely sister Maryam Aiyelero Mohammed thank you for being there for me, especially when I needed you most Jazaka Allah kheir. My Parents Mr. Aliyu Jimoh Mohammed and Mrs. Racheal Aliyu Mohammed for always being their Jazaka Allah kheir for your love and support during my difficult times. my sincere appreciation goes to my parent-in-law, Mr and Mrs Shehu Olaoti Jazaka Allah kheir for your support and encouragement during my period of study I am grateful. Lastly, my relatives, colleagues, friends and football teammates. You all made my academic journey worthwhile. Thank you

Abstrak

Pengesanan anomali teks berasaskan semantik adalah bidang penyelidikan yang menarik dan telah mendapat perhatian daripada komuniti perlombongan data. Pengesanan anomali teks mengenal pasti maklumat yang menyimpang daripada maklumat am yang terkandung dalam dokumen. Data teks dikaitkan dengan masalah kekaburan, keamatan tinggi, bersela dan perwakilan teks. Sekiranya cabaran ini tidak diselesaikan dengan baik, pengenalpastian anomali teks berasaskan semantik akan menjadi kurang tepat. Kajian ini mencadangkan Teknik Pengecualian Jujukan yang ditambah baik (ESET) untuk mengesan anomali teks berasaskan semantik dengan mencapai lima objektif: (1) untuk mengubahsuai Teknik Pengecualian Jujukan (SET) dalam memproses teks tidak berstruktur; (2) untuk mengoptimumkan Kesamaan Kosain bagi mengenal pasti data teks serupa dan tidak serupa; (3) untuk menghibridkan SET yang diubahsuai dengan Analisis Semantik Laten (LSA); (4) untuk mengintegrasikan algoritma Lesk dan Pemilihan Keutamaan bagi penyahtaksaan makna dan mengenal pasti bentuk kanonik teks; dan (5) untuk mewakili anomali teks berasaskan semantik menggunakan Logik Tertib Pertama (FOL) dan Graf Konsep Rangkaian (CNG). ESET melaksanakan pengesanan anomali teks dengan menggunakan Kesamaan Kosain yang dioptimumkan, menghibridkan LSA dengan SET yang diubahsuai, dan mengintegrasikannya dengan algoritma Penyahtaksaan Makna Perkataan khususnya Lesk dan Pemilihan Keutamaan. Kemudian, FOL dan CNG dicadangkan untuk mewakili anomali teks berasaskan semantik yang dikesan. Bagi menunjukkan ketersauran teknik tersebut, empat set data telah dipilih untuk diuji iaitu data NIPS, ENRON, blog Daily Koss, dan 20Newsgroups. Penilaian eksperimen menunjukkan ESET telah meningkatkan ketepatan pengesanan anomali teks berasaskan semantik daripada dokumen. Apabila dibandingkan dengan pengukuran sedia ada, keputusan eksperimen telah mengatasi kaedah penanda aras dengan skor F1 yang lebih baik daripada semua set data; Data NIPS 0.75, ENRON 0.82, blog Daily Koss 0.93 dan 20Newsgroups 0.97. Hasil yang dijana daripada ESET telah terbukti signifikan dan menyokong tanggapan yang semakin berkembang mengenai anomali teks berasaskan semantik dalam literatur yang sedia ada. Secara praktikal, kajian ini menyumbang kepada pemodelan topik dan pertautan konsep bagi tujuan menggambarkan maklumat, perkongsian pengetahuan dan mengoptimumkan pembuatan keputusan.

Kata Kunci: Kesamaan semantik, Anomali teks berasaskan semantik, Penyahtaksaan Makna Perkataan, Teknik Pengecualian Jujukan ditambah baik.

Abstract

The detection of semantic-based text anomaly is an interesting research area which has gained considerable attention from the data mining community. Text anomaly detection identifies deviating information from general information contained in documents. Text data are characterized by having problems related to ambiguity, high dimensionality, sparsity and text representation. If these challenges are not properly resolved, identifying semantic-based text anomaly will be less accurate. This study proposes an Enhanced Sequential Exception Technique (ESET) to detect semantic-based text anomaly by achieving five objectives: (1) to modify Sequential Exception Technique (SET) in processing unstructured text; (2) to optimize Cosine Similarity for identifying similar and dissimilar text data; (3) to hybridize modified SET with Latent Semantic Analysis (LSA); (4) to integrate Lesk and Selectional Preference algorithms for disambiguating senses and identifying text canonical form; and (5) to represent semantic-based text anomaly using First Order Logic (FOL) and Concept Network Graph (CNG). ESET performs text anomaly detection by employing optimized Cosine Similarity, hybridizing LSA with modified SET, and integrating it with Word Sense Disambiguation algorithms specifically Lesk and Selectional Preference. Then, FOL and CNG are proposed to represent the detected semantic-based text anomaly. To demonstrate the feasibility of the technique, four selected datasets namely NIPS data, ENRON, Daily Koss blog, and 20Newsgroups were experimented on. The experimental evaluation revealed that ESET has significantly improved the accuracy of detecting semantic-based text anomaly from documents. When compared with existing measures, the experimental results outperformed benchmarked methods with an improved F1-score from all datasets respectively; NIPS data 0.75, ENRON 0.82, Daily Koss blog 0.93 and 20Newsgroups 0.97. The results generated from ESET has proven to be significant and supported a growing notion of semantic-based text anomaly which is increasingly evident in existing literatures. Practically, this study contributes to topic modelling and concept coherence for the purpose of visualizing information, knowledge sharing and optimized decision making.

Keywords: Semantic similarity, Semantic-based text anomaly, Word Sense Disambiguation, Enhanced Sequential Exception Technique.

Table of Contents

Universiti Utara Malaysia

Universiti Utara Malaysia

Universiti Utara Malaysia

List of Tables

Universiti Utara Malaysia

List of Figures

Universiti Utara Malaysia

Universiti Utara Malaysia

List of Abbreviations

ANN	Artificial Neural Network
BCD	Block Coordinate Descent
CG	Conceptual Graphs
CGIF	Conceptual Graph Interchange Format
CNG	Concept Network Graph
ESET	Enhanced Sequential Exception Technique
FCA	Formal Concept Analysis
FCA-RS	Similarity measure proposed in Wang and Liu
FOL	First Order Logic
GSDPMM	Gibbs Sampling algorithm for Dirichlet Multinomial Mixture Model
GMM	Gaussian Mixture Model
HDP	Hierarchical Dirichlet Processing
HMM	Hidden Markov Model
k-NN	k-Nearest Neighbour
LCH	Leacock & Chodorow
LDA	Latent Dirichlet Allocation
LSI	Latent Semantic Indexing
LSA	Latent Semantic Analysis
MMR	Maximum Marginal Relevance
NED	News Event Detection
NER	Name Entity Recognition
NG	Network Graph
NIPS	Neural Information Processing Systems
NLP	Natural Language Processing
NMF	Non-Negative Matrix Factorization
NMI	Normalized Mutual Information
OLAP	Online Analytical Processing

Universiti Utara Malaysia

PCA	Principal Component Analysis
PMI-IR	Point wise Mutual Information using data collected
PLSA	Probabilistic Latent Semantic Analysis
PLSI	Probabilistic Latent Semantic Index
POI	Persons of Interest
POS	Part of Speech
RES	Resnik
SET	Sequential Exception Technique
SVD	Singular Value Decomposition
SVM	Support Vector Machine
WUP	Wu & Palmer
WSD	Word Sense Disambiguation

Universiti Utara Malaysia

Definition of Terms

Anomaly	observation which deviates so much from other observations as to arouse uncertainties that it was produced by an alternate mechanism
Algorithm	set of rules to be followed in solving a problem.
Semantic	meaning relating to words and phrases.
Corpus	collection of written texts, especially the entire works of an author or a body of writing on a subject.
Count vectorization	transformation of text into vector representations so that numeric machine learning approach such as counting can be done easily.
Disambiguation	removal of vagueness or ambiguity by making a context understandable or clear in meaning.
Dissimilarity	difference or variance
Cardinality	total count of elements present in a set or group, as a property of that group.
Measure	process of ascertaining the size or degree of an object.
Method	procedure or an approach of accomplishing a task.
Modification	the process of changing or adapting to improve an object
Network Graph	vertices or nodes that are connected by edges.
LAS	co-occurred terms found in corpus are captured using dimensionality reduction approach (SVD) on a term-by-document matrix *T* representing corpus
FOL	computational approach to knowledge representation following the language rules of grammatical representation.

Pruning	process of reducing the complexities of classifiers and hence improving its accuracy by reducing overfits.
SET	sequential exception technique recreates an approach in which unusual objects can be differentiated from series of like objects.
SVD	Singular Value decomposition is used to simplify or ease term vectorization in Text mining
Technique	way of carrying out an operation

Universiti Utara Malaysia

CHAPTER ONE

INTRODUCTION

1.1 Overview

This study presents an Enhanced Sequential Exception Technique (ESET) for semantic-based text anomaly detection. The study focuses on enhancing a technique that gives a better detection accuracy in identifying and representing semantic-based text anomalies in documents. To achieve this, chapter one was structured as thus; Section 1.2 briefly discuss the study research background. Section 1.3 states the research problem. Section 1.4 outlines the research question. Section 1.5 outlines the research objectives. Section 1.6 presents the research scope. Section 1.7 presents significance of the study and section 1.8 presents the organization of thesis.

Universiti Utara Malaysia

1.2 Research Background

Enhanced sequential exception technique was used in this study to detect semantic based text anomaly in documents. Hence, various unique methods have emerged over the years to satisfy the need of detecting semantic based text anomaly (Arning & Rakesh, 1996; Kamaruddin, 2011;. Kamaruddin et al., 2015; Kamaruddin, Hamdan, Bakar, & Mat Nor, 2012; Takahashi, 2011; Upadhyaya & Singh, 2012). With advancement in technology, the overload phenomenon of text document needs to be properly managed for knowledge sharing purposes and optimized decision making (Lee, et.al, 2017). Text information is one of the most valuable assets in the world today. Nonetheless, discovering meaningful knowledge from large volume of text document is tasking (Debortoli, Müller, Junglas, &

vom Brocke, 2016; Ramya, Venugopal, Iyengar, & Patnaik, 2016). This is due to the prevalent syntactic and semantic challenges present in text data. However, text data possess heterogeneity and high dimensionality. Dimensionality reduction is usually performed prior to applying various algorithms to avoid the effects of curse of dimensionality in text which leads to concentration of irrelevant text attribute, incomparable scores for different dimensionalities and hubness (i.e. objects occur more frequent in neighbour lists than others). Due to these reasons, the growth of research interest towards mining meaningful text through task like text clustering (Abdulsahib, 2015), text classification (Dang & Ahmad, 2014; Yoo & Yang, 2015) and text anomaly detection (Kamruzzaman, Haider, & Hasan, 2010; Mahapatra, Srivastava, & Srivastava, 2012; Kannan, Woo, Aggarwal, & Park, 2017) is on the increase. An in-depth review was performed in this study stemming from overflow of text information, methods used in identifying text semantics and how anomaly detection was employed in tackling existing issues related to detecting semantic based text anomaly for knowledge creation purposes.

Anomalous text are implicit knowledge that is distinctively different from general contextual idea (Hodge & Austin, 2004; Kamaruddin et al., 2015; Ramakrishnan Kannan, Woo, Aggarwal, & Park, 2017; Mahapatra et al., 2012). Detecting anomalous text refers to the task of identifying documents, or segments of text, that are unusual, rare or different from normal text (Guthrie, Guthrie, Allison, & Wilks, 2007). Text anomaly occur relatively infrequently, when they do occur, their consequences can be quite dramatic and often in a negative sense. In spite of its negative effects, anomaly detection has helped in attracting much attention in revealing meaning to disturbing events like the issues of information overload in text documents (Guthrie, Allison, & Wilks, 2007) and Socio-

political threats to national security (Terrorism) (Abouzakhar, Allison, & Guthrie, 2008). Consequently, a great deal of research has made it known that anomaly detection in text presents difficult challenges due to the nature of unstructured textual data (Chandola et al., 2009; Kamaruddin et al., 2012; Rockwell, 2003; Rozovskaya & Roth, 2013). This is due to text inconsistency and morphological drawbacks (Chandola, Banerjee, & Kumar, 2009; Mahapatra et al., 2012).

Many literatures have been proposed to effectively overcome the challenges of discovering semantic anomalies in an unstructured text data (Guthrie, Allison, & Wilks, 2007; Guthrie, 2008; Kumaraswamy & Shavlik, 2012; Mahapatra et al., 2012). Consequently, detecting text anomaly spans from different fields of study like Statistics, Text mining and Natural Language Processing (NLP) (Chandola et al., 2009; Kannan, Woo, Aggarwal, & Park, 2017). Statistical based approach is obliged to information generated by the parameters of data. Clustering and K-nearest neighbour are typical examples of the Distance-based approach which is proximity based. Apparently, clustering is slow because many cluster-based approaches rely on distance computation between text data with high linear dimensionality (Bhaduri, Matthews, & Giannella, 2011; Jain, 2010; Kannan et al., 2017; Miller & Myers, 2001). On the other hand, Classification-based approach offers promising results with methods like the Neural Networks (NN), Naïve Bayes and Support Vector Machine (SVM) (Ramakrishnan Kannan et al., 2017; Manevitz, 2001). Thus, these methods are all known to be applied when there is a distinguishable difference between anomalous classes and normal classes in documents.

However, these literatures made use of approaches that are computationally complex, requires computation of all pairwise distance between elements in data and uses data that

Universiti Utara Malaysia

requires training which is completely subjective to prior knowledge by users. There is a need to focus on methods that has linear complexity, less pairwise distance computation and can detect anomalies in text without training data or having prior knowledge.

1.3 Problem Statement

Researchers are beginning to attach significant importance to a better semantic-based text anomaly detection technique (Janz, Kędzia, & Piasecki, 2018). Anomaly-based approach finds data which are unusually different (either infrequent or frequent) (Mahapatra et al., 2012). It combines important properties of both the classification and clustering approach by finding labelled and unlabelled data to simplify the process of anomaly detection in text documents (Akoglu, Tong, & Koutra, 2014; Chandarana, 2015; Goldstein, Goldstein, & Uchida, 2016; Kim & Montague, 2017). This approach is considered viable because of its ability to detect text anomalies by examining their attributes, just like a similitude of human being seeing series of similar and dissimilar data (Guthrie, 2008; R. Kannan et al., 2017; Kumaraswamy & Shavlik, 2012; Mahapatra et al., 2012). Anomaly-based approach employing the Sequential Exception Technique (SET) by Arning & Rakesh (1996) and Zhang & Feng (2009) has proven to have a significant potential in detecting anomalies in categorical data such as log files from large databases. However, this research focuses mainly on detecting semantic-based text anomaly from documents. Therefore, the need to modify SET functions in processing text data is pertinent to this study. Moreover, text needs to be well pre-processed before it can be applied on modified SET for better semantic-based text anomaly detection in documents. Existing studies have studied many text pre-processing approaches such as Language Modelling approach, which has been identified to be computationally demanding (Classen, Boucher, & Heymans, 2011) and

the Hidden Markov Model (HMM), which is known to provide a significantly accurate result but most times is unable to capture relations in words (Ray & Craven, 2001). Apparently, Natural Language Processing (NLP) has been successfully used to overcome the problem of heterogeneity and text sparsity (Abdulsahib & Kamaruddin, 2015). But it was also noticed that the dissimilarity function in SET made use of variance and standard deviation which may not be as efficient in performing similarity/dissimilarity identification of term sequence in document (Deshpande, Vaze, Rathod, & Jarhad, 2014; Gabrilovich & Markovitch, 2007).

Term-based similarity measure is a viable measure for SET, because term-based similarity measures performed better in similarity/dissimilarity identification of term sequence. Nevertheless, not all term-based similarity measures are good for identifying term sequence. A typical example is the Jaccard distance, which considers mainly membership in terms and ignores term frequency (Gomaa, 2013; McInnes & Pedersen, 2013). Another example is the Euclidean similarity, its identification can be problematic if longer vectors have longer instances in documents (William Wei Song, Chenlu Lin, 2017). In operating with longer vectors many literatures made use of cosine similarity to identify text similarity in documents (Acree, Jansa, & Shoub, 2016; Deshpande et al., 2014; Gabrilovich & Markovitch, 2007). Imperatively, cosine similarity accounts for the ratio between words and discards words frequency by normalizing all text article into vectors to have uniform magnitude while maintaining the ratio between words. These attributes in cosine similarity has been used in achieving complex tasks like finding and grouping similar / dissimilar text documents. More so, it is pertinent to know that cosine measure incorporates more of linguistic structures using syntactic dependencies on textual data.

Universiti Utara Malaysia

Imperatively, both syntactic and semantic structure of text data are needed to detect semantic-based text anomaly in documents to consider a model that best analyse semantic-based text anomaly in documents.

According to Gomaa (2013) Pointwise Mutual Information - Information Retrieval (PMI-IR) computes similarity between pairs of words, which depends on text co-occurrence. The more frequent two words closely co-occur, the higher PMI-IR similarity score. While the Normalized Google Distance (NGD) is derived from the number of hits returned via Google search engine for a given set of keywords. Keywords with similar meanings in a natural language sense tend to be "close" in units of Google distance. NGD measures are solely dependent on Google search engine (Franzoni, 2017; Pradhan, Gyanchandani, & Wadhvani, 2015). A better semantic analysis approach is needed to resolve semantic related issues in text document. Existing studies on Latent Semantic Analysis (LSA) assumes that words that are semantically related will occur in similar pieces of text. A matrix containing word counts per word, phrase, sentence, paragraph and document is built from a large piece of text and a mathematical method called Singular Value Decomposition (SVD) is employed in LSA operations. It measures the effectiveness in producing more coherent term or concept models in text documents (Froud, Lachkar, & Ouatik, 2013; Henriksson, Moen, Skeppstedt, Daudaravičius, & Duneld, 2014; Rumshisky, 2008; Zhang, Xiao, Li, & Zhang, 2016). Again, it is sometimes possible that the absence and misuse of appropriate contextual meaning (ambiguities) are the root cause of unclear sentences in documents. This can be resolved by managing text information for easy semantics identification (Beltagy, Roller, Cheng, Erk, & Mooney, 2015; Faruqui, Tsvetkov, Rastogi, & Dyer, 2016; Nakov, 2013; Slimani, 2013).

Universiti Utara Malaysia

Resolving text synonyms and polysemous (ambiguity challenges) may be costly especially when there is a need to detect and analyse text semantics from huge numbers of documents (Abdulsahib, 2015; Abouzakhar, Allison, & Guthrie, 2008; Beltagy et al., 2015; Gahl et al., 2003; Kumaraswamsy & Shavlik, 2012; Kannan, Woo, Aggarwal, & Park, 2017 Mahapatra et al). A novel study performed by Kamaruddin, Hamdan, Bakar, & Mat Nor (2012) made use of embedded synonym identification with Conceptual Graph Interchange Format (CGIF) to match text semantic in documents. It was however noticed that, synonyms generated in every CGIF was costly, relies on word pairs in text documents "closed" synonyms and most time does not carter well for the predicate argument structure in sentences. Other existing research like (Abouzakhar et al., 2008; Montes-y-gómez, Gelbukh, & López-lópez, 2002) employed segmentation ambiguity to capture conjunctions in document paragraphs. This approach is computationally intensive, as it yields a low precision results compared to other approaches and may lead to character classification problems in text. A notion that greatly simplifies ambiguity task is vital. Zhang and Patrick (2005) employed text canonicalization to transfer texts of similar meaning into same surface text with a higher probability than those with different meaning. Bangalore et al. (2016) leveraged text canonicalization to fuse structured and unstructured text data to perform text pharmaco-vigilance information extraction and semantic identification. The obtained result from Bangalore et al. (2016) was significant and convincing enough to embrace text canonization in ambiguity related challenges especially in text data. Text canonization is important because, it simplifies the task of handling single meaning word representation for wide range of expression to disambiguate senses in document. More so, expressions can be easily related to that of Natural Language, which has been used in solving challenges like lexical ambiguity,

syntactic structure, syntactic ambiguity and POS (noun, pronoun, verb) resolution problem. Thus, there is a need for enhancement of ESET to tackle the text canonization problem.

A hybridized Enhanced Sequential Exception Technique with Word Sense Disambiguation (WSD) algorithm combining Lesk and Selectional preference algorithm to tackle text canonicalization problems was introduced in this study. In this approach, sense disambiguation was performed by leveraging both corpus and knowledge-based approach as well as employing non-hierarchical and hierarchical word relatedness for semantic-based text anomaly detection in documents. Another prevalent issue in this study is reliably represent the identified semantic-based anomalous text from huge numbers of text documents.

Many representation scheme has been employed in literatures such as Dependency graph, Conceptual graphs, Ontology and semantic kernel (Jiang, Zhang, Yang, & Xie, 2013; Kamaruddin et al., 2012; Poon & Domingos, 2010; Y. Wang, Ni, Sun, Tong, & Chen, 2011). These representation schemes are complex for relatively simple actions, also termed as a text-based on term frequency and approximation of lexical features. These representation schemes most times tends to ignore semantic content from large corpus (Kumaraswamy & Shavlik, 2012; Manevitz, 2001). Recently, literatures have emerged leveraging FOL for better semantic representation scheme in text data (Bruynooghe & Denecker, 2014; Garrette, Erk, & Mooney, 2014; Margaret Rouse, 2005). These literatures showed that First Order Logic (FOL) is promising in semantic representation and can be explored further for better results.

Universiti Utara Malaysia

In summary, one distinguishable difference of this study is the use of Enhanced Sequential Exceptional Technique (ESET) to match similar text and distinguish dissimilar ones. Compared to existing studies, Enhanced Sequential Exceptional Technique is tailored towards a simplified approach to lessen the complexity and improve the detection of semantic-based anomaly accuracy in text documents.

In summary, ESET is designed to tackle the following problems;

- Enhance SET to process text
- Optimize Cosine similarity with ESET to detect text anomalies
- Hybridize ESET with LSA to analyse semantics from anomalous text
- Canonize identified semantic based text anomalies using combined WSD algorithm namely Lesk and Selectional preference.
- Represent detected semantic-based text anomalies using FOL and CNG.

Universiti Utara Malaysia

1.4 Research Questions

The problem highlights the need to detect semantic based text anomalies. Hence, the following research questions are formulated:

1. How to modify Sequential Exception Technique (SET) functions in processing unstructured text data?
2. How to optimize cosine similarity /dissimilarity in identifying text anomalies from documents?
3. How to hybridize Enhanced SET with Latent Semantic Analysis (LSA) to detect semantic-based text anomaly?

4. How to integrate Lesk and Selectional preference algorithm by examining hierarchical relationship that identifies text ambiguity and analyse semantic-based text anomaly?
5. What is the most reliable representation scheme to capture semantic-based text anomaly?

1.5 Research Objectives

This study aims to enhance Sequential Exception Technique semantic based text anomaly detection. The following sub-objectives are postulated as follows to achieve the main objective

1. The study main objective is to enhance SET to detect semantic based text anomaly in text
2. To modify Sequential Exception Technique (SET) functions in processing unstructured text data
3. To optimize cosine similarity /dissimilarity in identifying text anomalies from documents
4. To hybridized Enhanced (SET) with Latent Semantic Analysis (LSA) that detects semantic-based text anomaly.
5. To integrate Lesk and Selectional preference algorithm by examining hierarchical relationship to simplify ambiguity by identifying text canonical form and analyses semantic-based text anomaly.
6. To represent semantic-based text anomaly using First Order Logic and Concept Network Graph.

1.6 Research Scope

Universiti Utara Malaysia

A narrower scope of this research was centred on ESET which was built on a modified SET (Optimized cosine functions with text-pre-processing techniques), LSA, WSD algorithms and FOL with Concept Network Graphs. This study adopts a quantitative research method leveraging the experimental design to critically study the research objectives using novel approaches for generating refined knowledge from corpus. The study data were basically sourced out from UCI Machine learning repository. As experiment is performed, smaller sample text data (sentences) were also evaluated concurrently. Nevertheless, ESET detections were limited to the identified frequent and infrequent text data. The generated output from ESET were used for decision and knowledge creation purposes. Conclusively, performance evaluation scores were benchmarked with other similar existing studies to measure accuracy of ESET. This was solely aimed at satisfying the study objectives as well as contributing to the body of knowledge in the field of text mining both practically and theoretically.

Universiti Utara Malaysia

1.7 Significance of the study

The research contributed to the theoretical body of knowledge by pointing out the need of enhancing Sequential Exception Technique for unstructured text data. This contribution supported a growing notion of semantic based anomalous text which is increasingly evident in existing literatures. However, it is empirically proven that frequent and infrequent text data may be contextually anomalous and as well convey meaningful ideas. The second contribution is the practical contribution. The study contributes to the body of knowledge practically by detecting side information and other forms of relevant text information. This is performed by incorporating different approaches such as similarity measures, topic model and Word sense disambiguating algorithms as a technique to detect

semantic based text anomalies. Lastly, a new representation scheme was presented by combining both FOL and Concept Network Graphs. This is aimed at improving understandability and interoperability of identified semantic based anomalous text data.

1.8 Organization of Thesis

Structurally, the study was divided into eight chapters, chapter one introduces the whole study by providing research introduction. It goes further to give the specific problems which the study addresses, pointing at the gaps in previous literatures which formulates the research questions and objectives of the study. Respectively, chapter one also provides some specific clarity on research scope.

Chapter Two reviews literatures that are relevant to the study. These literatures includes; Information overload, Text Mining, Canonical form with WSD, Text semantics representation, semantics analysis in text and anomaly detection. Also, from the reviews, these literatures formulate the research framework of the study.

Chapter three discusses the approach, strategy, algorithms and techniques employed in the study. It started by explaining the theoretical framework which guides the study, it then goes further to explain the research design of the study. At the end of this chapter, procedures and techniques of data evaluation was also discussed. In chapter four, the explored techniques with some parts of results of the research were discussed. Generated results from this chapter were also used to answer some parts of the research questions highlighted in chapter one. Different phases of ESET were discussed for chapter five, six and seven to answer the research questions of the study. These chapters provide an in-

Universiti Utara Malaysia

depth explanation of the phases. Chapter Eight provided solutions with highlights on the implication of findings of the study for future researchers with conclusion of study.

Universiti Utara Malaysia

CHAPTER TWO

LITERATURE REVIEW

2.1 Literature Background

An in-depth review was made in this research, stemming from overflow of unstructured text information to methods used in detecting and representing semantic-based text anomalous. Detecting anomalous text refers to the task of identifying documents, or segments of text, that are frequent or infrequent text (Guthrie et al., 2007; Mahapatra et al., 2012). It is imperative that when text anomaly occurs, their consequences can be quite dramatic and often in a negative sense. In spite of its negative effects, text anomaly detection has helped in attracting much attention in revealing meaning to disturbing events like the issues of information overload in text documents (Guthrie, Allison, & Wilks, 2007) ,Socio-political threats to national security (Terrorism) (Abouzakhar et al., 2008) and other interesting issues relating to identifying meaning in textual documents (Cambria & Melfi, 2015; Gilad Katz, Yuval Elovici, 2014; Kamaruddin et al., 2012; Mahapatra et al., 2012). A systematic literature review was made on existing related studies in this chapter for detailed understanding of the research objectives.

2.2 Unstructured Text Information

The access to constant flow of information is a golden opportunity and as well a big challenge for organizational control and information management. However, to gain easy access to meaningful information, core ideas must be discovered. In fact, a recent study showed that 80% of company's information is contained in text documents (Debortoli et

al., 2016; Ramya et al., 2016). The process of discovering and identifying useful text information can be well explained by exploring the techniques in text mining. But before text can be mined, text as a data needs to be understood. Basically, text is either structured or unstructured in nature. This study focuses on unstructured text data. Text information that does not have a well-defined systematic structure is said to be unstructured (Poon & Domingos, 2010).

Text are typical strings of character represented as units of meaningful combination of characters in Natural Language. These characters are basic textual units that are regarded as word, phrases or sentences (Ngai et al., 2016). According to Jurafsky & Martin, (2000) the creation of meaningful text representations involves a wide range of knowledge-source. Therefore, it is believed that knowledge discovery from textual databases employing text mining task has a higher commercial potential in the field of Artificial Intelligence (Katariya & Chaudhari, 2015). However, the study makes use of a novel technique that identifies knowledge through semantic-based text anomaly detection as will be explained and reviewed in the subsequent sections.

Universiti Utara Malaysia

2.3 Text Mining

Text mining task can be challenging especially when dealing inherently with unstructured textual data. There are many text mining tasks that can be applied on unstructured textual data. Text mining brings upon the contributions of different text analytical components and knowledge input from disciplines like Artificial intelligence, Statistics, Computer science and Machine Learning. These results in decisions affecting fields like information retrieval, natural language, web mining, classification and clustering (Aggarwal & Zhai,

2012; H, M, & Science, 2015). This research focuses more on machine learning approach as a component in Artificial Intelligence such as the supervised (classification) and unsupervised (clustering and anomaly detection) machine learning approach (Patel & Soni, 2012) in achieving the research objectives of the study.

Table 2.1.

Text mining approach.

Methods and Authors	Advantages	Disadvantages
Supervised Classification based method SVM, SVM Cichosz, (2018); Janz et al., (2018); Kim & Montague, (2017) Manevitz & Yousef (2000) Abdul –Jaleel et al, (2004); Manevitz Yousef (2000) Srivastava et al. (2006)	parameter need to be optimized it is best applied when assigning instances to an appropriate type of a known data type	Time consuming for high dimensional data Needs predefined deviating category High computation cost
Unsupervised Clustering based Graph based, NN Arning et al. (1996) Kamaruddin et al., (2015) Osmar, et al., (2014) Cichosz, (2018); Srivastava & Zane-Ullman (2005); Zhang et al. (2004) Akarsu, et al.(2013)	It requires to satisfy scalability, ability to deal with attribution of different types and ability to deal with noise and anomalies. No prior data distribution is needed Ability to adopt to new cases Able to process large data in linear time	Difficulty in identifying optimal parameter for the distance computation Too sensitive to the arrangement of input

Table 2.1 shows text anomaly detection techniques using the unsupervised machine an in-depth research must be done on other text mining tasks to have a clearer understanding on what they entail in detecting anomaly in text.

a. Text Classification: Text categorization, topic classification and topic spotting are the process of assigning text documents into various categories. Text classification involves assigning a predefined categories of text documents such as web pages, news stories and technical reports. These categories of text document are most times pertinence or topics (W. Zhang, Tang, & Yoshida, 2015). However, the notion and idea of classification is very general in text mining. Its application goes beyond information retrieval (Christopher et al 2008). A study made use of the classification approach to build an application that identifies Fraudulent refunds(Issa & Vasarhelyi, 2011). Another research made use of the classification approach to detect anomalous text data in the value of domain knowledge. The study shows that the domain-specific features are more predictive and that the relational learning methods exhibit superior performance (Kumaraswamy & Shavlik, 2012). Text classification is best applied when assigning instances to an appropriate type of a known data type (supervised data). The goal of text categorization is to classify documents into a fixed number of predefined categories, where each documents can be in a multiple form, can be classified as exactly one or have no category at all (Joachims, 1998). The main issue with this approach is that it has no available accurate labels for various normal classes. It also assigns label to each test instances, this becomes a problem when meaningful anomaly score is desired for this test instance that has become a subject to Classification based technique (Upadhyaya & Singh, 2012).

b. Text Clustering: refers to the process of finding groups of similar text elements, which are collected together for a specific purpose in unstructured formal documents. It details with finding a structure in collection of unlabelled data (Brody, 2005; Kumar,

2012). More so, the definition of documents being similar or dissimilar is not always ambiguous, which varies with the actual problem or domain setting. For instance, the process of clustering research papers into two documents would be regarded as same if they share similar thematic topics (Huang, 2008). Clustering text data has an exceptional feature of digesting and generalizing good amount of text information in documents. Document clustering is a vital technological approach in recent years due to its vital techniques in text mining. A research by Akarsu, Bayram, Slisko, & Corona Cruz, (2013) cluster sentence level using fuzzy relational text clustering algorithm to identify overlapping semantically related text clusters. Another research by Mahapatra et al (2012) made use of clustering to identify contextual anomaly in documents using the LDA to model topics. Both research results were significant compared with other benchmarked approaches used in achieving similar objectives.

The goal of text clustering is to identify the intrinsic group in a set of unlabelled data. There is no absolute criterion to decide what constitute best clustering practices. Consequently, it is the user who must provide this necessary criterion in a way that the result of clustering will suits their needs. Clustering algorithm is required to satisfy scalability, ability to deal with attribution of different types and ability to deal with noise. However, it is pertinent in this research to identify and detect useful information in corpus by detecting anomaly in text.

c. Text Anomaly Detection: Arning & Rakesh, (1996), stated that anomaly detection is a task that is similitude to human being seeing series of similar data. This feature allows anomaly detection to operate with ease by proffering suitable solution in detecting anomaly in text. It does this by examining the features of objects, either by

unsupervised machine learning or supervised machine learning approach of data (Schlesinger & Hlavác, 2011). Text anomaly detection method performs its task by examining the main features of an instance that does not conform or that are anomalous from other characteristics or features in a data set. It can also be employed in identifying rare patterns in text. These has led its application to many domain for various purposes such as Topic Modelling (Zhang et al., 2009), Text plagiarism detection (Oberreuter & Velásquez, 2013) Identifying anomaly in news (Montes-y-gómez et al., 2002) Extracting Information in medical text documents (Meystre et al, 2008) and Novelty detection in Business Blogs (Liang, Tsai, & Kwee, 2009).

Text Anomaly detection can be performed in different levels namely; Topic level (Allan, Carbonell, & Doddington, 1998), Event level (Brants, T., Chen, F., & Farahat, 2003), sentence level (Kamaruddin, Hamdan, & Bakar, 2007; Li, Member, & Croft, Head, 2006), Document level (Kamaruddin et al., 2015; Karkali, Rousseau et al., 2014) and Word level (Gabrilovich, Evgeniy, 2005). However, the choice of which level to use is dependent on the research objectives of the entire study. Furthermore, it is important to know if the selected level can be able to represent conveyed meaning explicitly in text documents. It is evident from previous research (Abdulsahib & Kamaruddin, 2015; Almarimi & Andrejková, 2016; Bernotas, Karklius, Laurutis, & Slotkiene, 2007; Brants, Chen, & Farahat, 2003; Cichosz, 2018; Gabrilovich, Evgeniy, 2005; Kamaruddin et al., 2015; R. Kannan et al., 2017; Karkali et al., 2014; Li et al., 2006) that core concepts of text documents emanates from sentences. Sentences tends to express vital and unique concepts from its terms.

2.4 Text Anomaly

Text anomaly are most times referred to as anomaly detection in text, text novelty and exception mining (Jacquenet & Largeron, 2009). However, various domain of application has inevitably raised issues in defining text anomalies. Some related statistical studies defined anomaly as an inconsistent subset of observation. Hence, for unstructured text, the subjective and consistent nature is more glaring since sentences that may be familiar to some seem different from others (Hodge & Austin, 2004). Naturally the goals for text anomaly is to detect useful data. Many researchers have adopted different approaches of mining anomalous text data in the field of text mining (Guthrie et al., 2007; Ramakrishnan Kannan et al., 2017; Kumaraswamy & Shavlik, 2012; Mahapatra et al., 2012).

Anomaly detection simply involves a learning task which most times use the unsupervised approach in creating a predictive historical data model to detect anomalous instances in new text data (Cichosz, 2018). In a broader sense, there are basically three types of anomalies. These are the collective anomalies, point anomalies and contextual anomalies.

Universiti Utara Malaysia

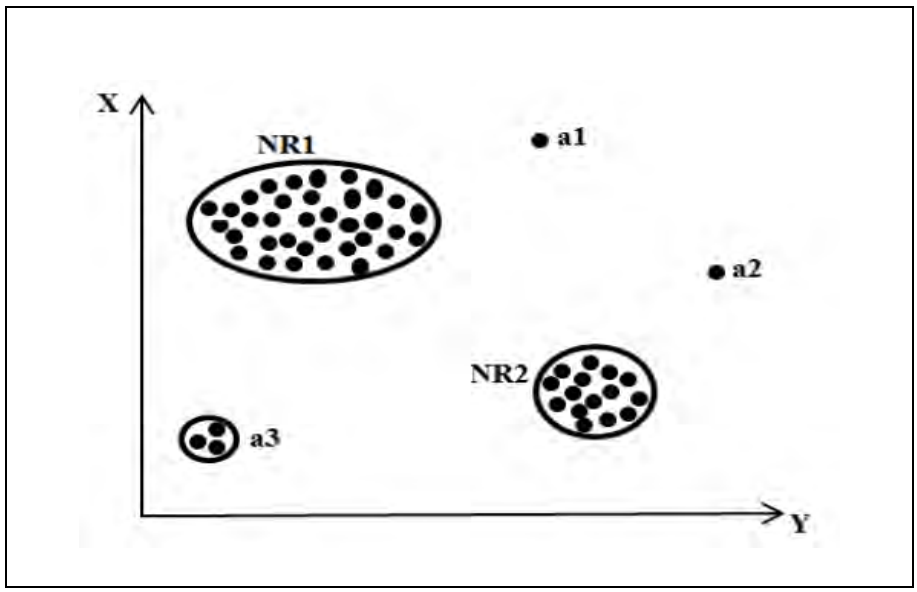

Figure 2. 1. Anomaly in X & Y Plane
Source: (Chandarana, 2015; Chandola et al., 2009)

Figure 2.1. described the three types of anomaly using the X & Y plane to illustrate the characteristics of point, collective and contextual anomaly. *a1, a2 & a3* depicts regions that are far away from other data and individual data instances are inconsistent with respect to the remaining set of data, then instances are labelled point anomaly. Contextual anomaly is defined by the attributes of its instance. While collective anomaly as shown in Figure 2.1 above *NR1 & NR2* are seen to be the normal regions of data sets. Since number of observation lies in these two regions (Pawar, 2015). Anomaly detection has been used to solve different challenges from exiting research using different approaches, such as the clustering or graph-based detection approach (Akoglu et al., 2014; Wang et al., 2011).

According to Adler-Golden, (2009) anomaly detection seek to detect interesting objects that are uniquely distinguishable from other data. Its abnormal or non-conforming pattern in different domain are referred to as discordant, anomaly, observation, anomaly, exception or contaminants. It main objective is to recognize a group of instances that are infrequent in a given data (Pawar, 2015). It is extensively used in discovering fraudulent

Universiti Utara Malaysia

credit cards, socio-political threats to national security (Terrorism) (Abouzakhar et al., 2008) in the health care system, intrusion detection in cyber security. Hence, its major issue with textual data is size, contextual meaning of data, curse dimensionality and scalability. Most learning techniques find it difficult to deal with all these issues (Manevitz, 2001).

The study will focus more on how anomaly can be leveraged to uncover text semantics. But prior to this, it is necessary to note the importance of text semantics and how anomaly can be employed in detecting text meaning from all levels.

2.4.1 Levels of Text Anomaly Detection

Text anomaly detection has been performed in different levels: the document level (Liang et al., 2009; Yang et al, 2002) the topic level (Yang et al., 2002), the event level (Allan et al., 1998; Brants et al., 2003; Yang et al., 2002) the sentence level (Allan et al., 1998; Breja, 2015; Cammert, et al., n.d.; Li et al., 2006; Otterbacher & Radev, 2006; Takahashi, 2011; Tsai, 2007; Yuhanis, 2015) and the word level. In this section, related literatures on these levels are briefly reviewed.

i. Topic Level Anomalies: Topic level anomaly detection refers to anomaly detection that focuses on certain topic predefined by users such as a query. The need to identify topic level anomalies is highlighted by Yang et al. (2002) where they reason that the same set of keywords is usually used in same topics. In their work, they performed the topic level anomaly detection by proposing a two-step process. The first step involves the classification of documents into predefined broad topics. The second step involves the first story detection in each of the defined topic.

Zhang et al. (2002) also worked on topic-level anomaly detection in which they performed adaptive filtering using statistical models to find relevancy and redundancy. In this work, document streams are filtered in two stages. The first stage is to find documents that have the same topic specified by user and the second stage is to find documents that contains new information compared to previously seen documents. Several researchers performed anomaly detection on subtopic level (Zhai et al. 2003; Dai & Srihari 2005). Zhai et al. (2003) considers subtopic level anomaly detection as a subtopic retrieval problem of retrieving as many documents as possible that cover different subtopics.

The documents were represented using Language Models and Maximal Marginal Relevance (MMR) technique was used to identify deviating subtopics. They concluded that even though both relevance and redundancy is important for subtopic retrieval, the relevancy is a more prevailing element. Dai and Srihari (2005) assumes each topic in a query to have different subtopics. Therefore, the retrieval and ranking of relevant documents was done with maximum coverage of subtopics but with minimum redundancy.

Universiti Utara Malaysia

They concluded that extracting and classifying documents according to subtopics allows users to specify a subtopic threshold and to adjust redundancy threshold.

ii. Event Level Anomalies: Topic level and event level are related in some sense. Yang et al. (2002) differentiated topic from event by providing some definitions as follows. Event are something that has happened somewhere at a certain time whereas topic are general events. For example, airplane accidents are topic while

a TWA-800 crash is an event. Using this definition, they modelled topic to comprise several events and events to have a set of documents. A topic-conditioned feature weights was proposed in which the weights are used in the calculation of event level deviating documents. Thus, a deviating event must be relevant to a topic and should discuss a new event. In other words, researchers identify event as "narrowly defined topic" (Allan et al. 1998; Yang et al. 2002). As a result, events in these works are represented as sets of related documents. (Allan Collins, Larkin, & Newman, 2007) and Allan (2004) investigated event level anomalies by proposing a multi-stage new event detection (NED) system.

In this work, the stories were classified into categories and NED was implemented in each category. Another research by Atefeh & Khreich, (2015) made a survey on techniques for event detection in twitter microblogs. Event detection is important to identifying key information relating to scenario over a period.

Universiti Utara Malaysia

iii. Document Level Anomalies: Document level anomaly detection aims to find relevant documents given a stream of documents. Most of the work in this level focuses on comparing a new document to all the documents in the past. Gabrilovich et al. (2004), presented algorithms that can identify deviating documents by analysing a series of newsfeeds articles and comparing it to the articles that have been read by the user. The technique was developed to analyse inter and intra document dynamics.

Intra-documents are concerned with how information evolve within individual article and the Inter-documents is concerned with how information evolve over time from article to article. However, their work requires pre-categorized

documents, i.e. the work assumes that the documents are already grouped into categories according to their contents. Yang et al. (2002) proposed a document level anomaly detection by using predicted topic for a certain document to evaluate the novelty of a new document. A novel approach for Novelty Detection of Web Documents. The novelty detection aims to build automatic systems which are capable to ignore old stories, essays, reports and articles already read or known, and notify the users of such systems about any new stories, essays, reports and articles (Breja, 2015). An algorithm was presented in mining text documents to discover anomalies by proposing a text mining system that is able to detect sentence anomalies from a collection of financial documents.

The system implements a dissimilarity function to compare sentences represented as graphs. Evaluation on the system revolves around experiments using financial statements of a bank. The findings provide valid evidence that the system can identify deviating sentences occurring in the documents. The detected anomalies can be beneficial for the authorities in order to improve their business decisions Kamaruddin et al. (2015).

iv. Sentence Level Anomalies: Sentence level anomaly detection refers to the task of searching for relevant and novel sentences given a query and a stream of relevant sentences or documents. A considerable amount of literature has been published on sentence level anomaly detection in the Text Retrieval Conferences (TREC) Novelty Track (Soboroff & Harman 2003). In (Allan et al. 2003) the sentence level anomaly detection was divided into two subsequent tasks i.e. finding relevant

sentences from the given documents and finding novel sentences among the identified relevant sentences. Besides identifying the coverage of a certain topic in the examined sentences, emphasis is also given in determining whether new information about a certain topic is presented in the sentences being analysed.

The results are list of relevant sentences and from these relevant sentences, novel sentence were marked as anomalies. Li and Croft (2005) performed novelty detection on sentence level patterns and argued that sentence patterns may be more relevant than individual words because sentence pattern consists query words, specific user requested entities and important phrases. This study extracted patterns in sentences which includes both query words and answer types. The former are possible answers for the query that might be present in the sentences. The study further identifies anomalies by picking out novel sentences that have new unseen answers. In 2006, Li and Croft enhanced their method by exploiting on how to manipulate the named entities and identifying additional information pattern that may be useful. Gamon (2006) represented sentences as connected graph without using linguistic analysis.

In this work, the term in sentences are represented as vertices and point wise common information between terms are represented as edges of the graph. Otterbacher and Radev (2006) proposed sentence level anomaly detection with fact-based relevancy detection. Given a user request on a certain topic, and the documents relevant to that topic have been identified, they developed a method first to identify sentences that provide the answers for the user request, second to

gauge previously unseen answers. Zhang and Tsai (2009) explored named entity recognition (NER) and part of speech (POS) tagging to perform sentence-level anomaly detection. Two sentences are compared to find the overlap of entities in the sentences. The overlap of other meaningful words is also captured in the residual-word novelty score process. Both the entity overlaps score and the residual-word overlap score were given weights before it is combined into a linear sum. A similar method was explored by Ng et al. (2007) where in addition to using NER and POS, synonyms for the entities and parts of speech is generated as well using WordNet. WordNet is an English lexical database of concepts and relations. They later determined the similarity between sentences using metrics such as Unique Comparison and Importance Values. Kamaruddin et al. (2015) explored the detection of sentence anomalies from a large collection of bank financial reports or documents. The system implements a dissimilarity function to compare sentences represented as graphs.

Universiti Utara Malaysia

v. Word Level Anomalies: Word level anomaly detection was explored in Gabrilovich et al. (2004). Here, the focus is on individual words, in which statistics regarding word occurrence across multiple documents is captured to identify similarity and dissimilarity between them.

Besides, modelling bag of words, named entities are extracted as well to identify common entities that are present in most news documents such as people, organizations and geographic locations. Although the authors claim that their method are comparable in performance to methods that manipulate documents in

other levels, most researchers argue that modelling individual words failed to capture the meaning represented by the documents. Text Anomaly can be detected in various level according to existing study. These levels of text anomaly are somewhat related in the sense that they are all text data. However, every level of text anomaly detection can be tailored towards achieving different objectives. The outlined levels of text anomaly detection are shown in Figure 2.2. below.

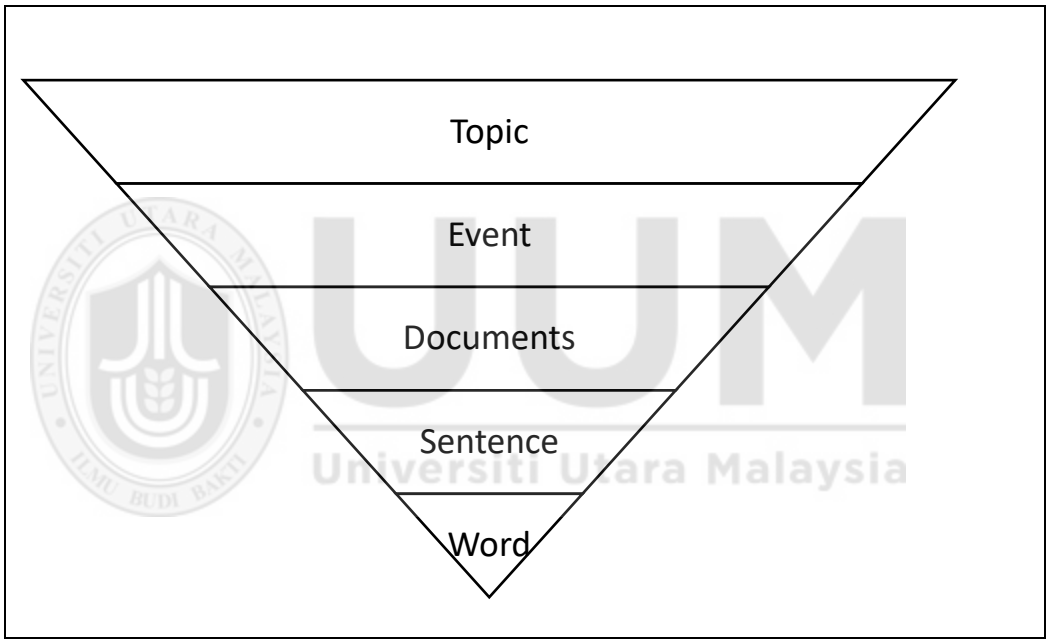

Figure 2. 2. Levels of Text Anomaly Detection

Figure 2.2 illustrates the hierarchical order of text anomaly detection starting from Topic to word detection. These levels of text anomaly detection have different roles to play and methods of detection may differ as well. To have an in-depth look on this, literatures were reviewed to understand text anomaly detection more.

2.4.2 Current work in Text Anomaly Detection and Text Semantics

A significant number of researchers have made an interesting contribution in identifying

text semantics using anomaly detection as an approach. A promising work by Mahapatra et al., (2012), used information to further anomaly detection algorithm in analysing semantic text data by considering divergence in statistical pattern seen from general semantic expectations. Experimental results revealed that their algorithm performed as expected, which could potentially minimize false positive rates in existing anomaly detection systems. Another interesting work by Kumaraswamy & Shavlik, (2012), postulated the hypo-study that, domain-specific features are more important than the linguistic features using the classical anomaly detection algorithms. First-order predicate logic was employed to demonstrate the effectiveness of domain knowledge in two different domains. Experimental result showed that the domain specific features are more predictive and the relationship learning methods exhibit better operational performance.

Kamaruddin, Hamdan, Bakar, & Mat Nor, (2012) captured semantic in financial text documents using Conceptual Graphs (CG) and Conceptual Graphs Interchange Format (CGIF). These researches have successfully embedded concept synonyms into CG and CGIF using dissimilarity functions. The dissimilarity score produced in these researches are strongly correlated with human evaluation of sentence similarity.

Universiti Utara Malaysia

2.4.3 Types of Anomaly Detection

Anomaly detection techniques in the context of textual domain is aimed at uncovering novel and interesting ideas, topics, discourse in document. Detected anomalous text data in corpus is most times represented as document to word co-occurrence matrix, which is usually sparse with high dimensionality. Curse dimensionality in textual data is a major hindrance in text anomaly detection. According to (Balbi, 2010), curse of dimensionality

is defined as “dimensions increase, the space in which individuals are represented becomes smaller, and it is difficult to find interesting relations between variables”. Hence, it is difficult to cluster data that have similar attributes. As in the case of unstructured textual data, the study considers this kind of data as text analysis issues like noise, semantic and syntax problem (Abouzakhar et al. 2008). From many literatures, it was noticed that there is no genuine knowledge on the type of text data that forms semantic anomalous text data in large documents. This will involve a lot of trial and errors using different approaches, techniques and algorithms. Many techniques have different approaches to find semantic in anomalous text data. Depending on the evaluation technique used in identifying anomaly in text semantic, many researches based their techniques on the following; Statistical, Distance and anomaly techniques.

Table 2.2.

Comparison anomaly detection approaches

Approach	Methods	Advantages	Disadvantages
Statistical (M. J. Denny & A. Spirling, 2018); Bertoldi, Cettolo, & Federico, (2010); Foltz, (1996); Pawar, (2015); G. Zhou, Zhao, Liu, & Cai, (2011)	Histogram based Kernel function based Regression model Gaussian Model	It is associated with a confidence interval, which can be an additional information while making decision regarding any test instance. It can operate in an unsupervised setting without any need for labelled training data.	It is obliged to the information generated by the parameters of data

Distance (Franzoni, 2017); Han, (2014); Tan, Steinbach, & Kumar, (2006); Sugiyama & Borgwardt, (2013); Bhaduri et al., (2011)	K-Nerest Neighbour Density based Grid Linearization Disk Block Access	They are data-driven and do not make assumptions regarding distribution of data.	It requires computation of all pairwise distance between elements in a data. Its time complexity leads to runtime problems on massive data
Anomaly Detection	Supervised K-Nearest Neighbour	Better dealings with text data without any systematic trends.	It requires thresholds for minimum size and distance.
(Arning & Rakesh, 1996; Guthrie, 2008; Kamaruddin et al., 2015; Montes-y-gómez et al., 2002; Zhang & Feng, 2009 Kannan et al., 2017)	Naïve Bayes Decision Tree Support Vector Machine Unsupervised Association based Clustering based • OLAP • SET	It also improves the accuracy of language model. outputs can be labelled, and it is easy to collect co-occurrence in text data which makes the acquisition of large amount of semantic knowledge from documents becomes feasible	Features are usually decided by experts and the frequent change in technology affects method

a. Distance-based approach

In Distance based method, distance between each data points are calculated. If the result of distance is over some predefined edge, target instance is considered anomaly (Pawar, 2015). Several distance base algorithms have been proposed recently to detect anomalous data. These algorithms are based on calculating the dimensional distance of different points using all available features, such as calculating the densities of local neighbourhoods of data. These algorithms are K-nearest neighbour, indexed based, Nested loop based and Cell based, proximity based and employs techniques like

partitioning, local & global technique to simplify detection processes.

Abouzakhar et al., (2008) introduced the problem of anomalous text detection in corpus by automatically extracting and evaluating linguistic feature patterns from suspicious information in Arabic documents. Three measures of dissimilarity function were adopted (Chebychev, City block and Cosine similarity) with an encouraging result. This aimed at improving Arabic language analysis and processing in NLP community. The dissimilarity measures distance of each segment in paragraphs from every other segment in document. This is frequently repeated for at least 30 times for accuracy to be achieved. Repetition of measured distance may be time consuming and tedious. So, further research on other literatures that leverage the use of (dissimilarity technique). Challenges of this approach includes retrieving of text data from different data source and its scalability. Since it requires computation of all pairwise distance between elements in a data. Its time complexity leads to runtime problems on massive data (Han, 2014; Sugiyama & Borgwardt, 2013).

Universiti Utara Malaysia

b. Statistical-based approach

Statistical methods observe the attributes of data by analysing and evaluating some important variables over a given period. By this observation of data, a standard model is formed, which follows a probability model (Pawar, 2015). The purpose of this approach is to detect anomalous data that is different from the standard model. Statistical based method are based on assumptions that is “Normal data instances occur in high likelihood regions of a model, while anomalies occur in the low likelihood regions of the model”(Chandola et al., 2009).

Basically, there are two types of statistical methods, which are Parametric and Nonparametric. They both have been applied to fit in statistical models. The parametric technique is assumed that normal data are generated by parametric distributions with parameters (Pawar, 2015). While Non-parametric statistical techniques model structure is not defined, but it is derived from a given data. Compared to parametric techniques, fewer norms are made like density smoothness. A typical example of a Non-parametric technique is the histogram (Chandola et al., 2009; Pawar, 2015). The major limitation of this method is that they work best with univariate data. Moreover, statistical method is obliged to the information generated by the parameters of data (Pawar, 2015). Many researchers made use of the statistical based method (Bertoldi et al., 2010; Foltz, 1996; Pawar, 2015; G. Zhou et al., 2011). Guthrie, (2008) made use of statistical method, this research was based on a supervised detection of huge segment of sentences. Different approaches were conducted to investigate whether it is possible to automatically identify anomalous segments of sentences like detecting subversive text in newswire. Some of the limitations of this research include tailoring of document flow towards single documents, also simple chunk of polysemous words and phrases were ignored. In addition to this it was also noticed that data require training which is completely subjective to prior knowledge by users. Due to this fact, information like data dispersal may not be known at times. In anomaly detection, statistical method does not promise to find all anomalous data where no specific test was carried out.

c. Anomaly based method

As earlier stated, this method has received much attention from many researchers over the years. Studies have affirmed a positive performance with the use of unsupervised learning

in anomaly detection method. According to (Kamaruddin, Hamdan, Bakar, & Mat Nor, 2009; 2012) semantic in financial text documents were captured using Conceptual Graphs (CG) and Conceptual Graphs Interchange Format (CGIF). These literatures have successfully embedded concept synonyms into CG and CGIF using dissimilarity functions. The dissimilarity score produced in these researches are strongly correlated with human evaluation of sentence similarity. However, (Kamaruddin, Hamdan, Bakar, & Mat Nor, 2009; 2012) focused on the collection of financial text document based on parser. Result of text document parsed is sentence structured. This implies that CG generator implements parsing by linking formal grammatical system to produce syntactical relations between words. The accuracy of CG depends more on syntactic existence in linguistic resources than semantic existence in document. Cambria & Melfi (2015) used Least Absolute Anomaly (LAD) in semantic anomaly detection to improve analogical reasoning in Affective space. Data used in the research was to test the combined detection of opinion targets and the polarity associated with it. Although the computational efficiency was a little complex and difficult. Almarimi & Andrejková (2016) developed an algorithm to cover text anomalies using histograms of mapped time sequences that was modified by a kernel smoothing function. Illustrations of results for Six English and Arabic texts respectively showed anomalies in artificial combined texts. Cichosz, (2018) addresses the possible promising text anomaly detection of Polish Internet discussion forum devoted to psychoactive substances received from home-grown plants, such as hashish or marijuana. This study made use of word embeddings approach (Global Vectors) combined with two unsupervised anomaly detection algorithms, based on one-class SVM classification and based on dissimilarity to k-medoids clusters. The study made a significant finding in detecting post related to psychoactive substances from the Polish

internet forum.

Considering the divergent application and limitations of past literatures in preceding paragraphs, anomaly-based method seems more significant and useful to detect semantic-based text anomaly. It may be a supervised or an unsupervised approach. This research is interested in using the unsupervised approach in anomaly detection which are OLAP (Online Analytic Processing) and SET algorithms. SET compares element in a set using functions like the dissimilarity, similarity, cardinality and the smoothing factor to identify anomalies. While OLAP is employed in analysing Business data in search of Business Intelligence used for discovery driven exploration for computational efficiency. It is implemented using cells stored in tables in a relational database. In contrast to this, the study is aimed at analysing unstructured text data not limited to BI alone (Furtado, Nadal, Peralta, Djedaini, & Marcel, 2015; Pawar, 2015).

Universiti Utara Malaysia

2.5 Sequential Exception Technique (SET)

This technique recreates an approach in which unusual objects can be differentiated from series of like objects. The next obvious question is why is SET important? Its applicability can be employed in different domains such as Bio-medicine, Text retrieval Network intrusion detection and Video resolution (Capurro et al., 2016; Héas, Drémeau, & Herzet, 2016).

Detection of text and optimal mapping in document employed dissimilarity function as a method to recognize plagiarized text (Stamatatos, 2009). Jimenez, Becerra, and Gelbukh, (2012) made use of delicate cardinality to provide an adaptable examination between words in text. Arning & Rakesh, (1996) used the Sequential Exception Technique (SET)

for categorical data of log files from large data. This technique was also used and enhanced by Z. Zhang & Feng, (2009) which was also applied on categorical data.

a) Dissimilarity function d: $\mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is usually understood to measure distance between points. Dissimilarity functions increase the more apart two points gets.

$$\text{A dissimilarity function } D: P(I) \rightarrow \mathbb{R}+0 \tag{2.1}$$

be the variance of numbers in a set. Z. Zhang & Feng, (2009) used dissimilarity function to detect anomalies in large database. Three algorithms were employed, and all these algorithms made use of the dissimilarity function. These algorithms are; the time proportional to the square of the data length, the second is time proportional to the square of the number of distinct data values and the third algorithm defines the anomaly factor. Bustince et al. (2011) focused on dissimilarity functions and how it can be analysed to generate aggregation functions. These functions can be of important to detecting anomalies in data.

Universiti Utara Malaysia

In the context of text mining distance with dimensions representing features of objects are leveraged to the fullest to check for similarity and dissimilarity. If distance is small, it means high degree of similarity where large distance means low degree of similarity. The dissimilarity is subjective and is highly dependent on the domain and application for instance two objects are similar because of colour or size. Proper care should be taken when measuring distance across dimensions or features that are not related. Relative values of each element must be normalized, or one feature could end up dominating the distance. Dissimilarity are measured in the range *0 to 1 [0, 1]*. Two main consideration about dissimilarity:

Dissimilarity = 1 if $X = Y$ (Where X, Y are two objects) Dissimilarity = 0 if $X \neq Y$ (2.2)

This study tends to make use of the sequential exception technique succinctly on Text data keys as well as values to detect anomalous features in data. Arning & Rakesh, (1996) made use of variance to easily detect anomalous data, other dissimilarity functions such as Cosine, Eugene Euclidean, Jaccard and Manhanttan can also be used as a function or measure for similarity and dissimilarity measurement (Henriksson et al., 2014; Huang, 2008; Szmeja, Ganzha, Paprzycki, & Pawłowski, 2018).

b. Cardinality function is used to count the occurrence of data

Cardinality function C*:*
$P(I) \rightarrow \mathbb{R}+0;$ *with* $I1 \subset I2 \rightarrow C(I1) < C(I2)$ *for all* $I1, I2$ (2.2)

c. Smoothing function it helps to remove noise from a target set. In most cases function $C(I - Ij)$ may return $|I - Ij|$, the number of items in the set $I - IJ$. This function may also be defined by formula $C(I - Ij) = 1 \div (|Ij| + 1)$ (Arning & Rakesh, 1996). Therefore, $SF(Ij) = C(I - Ij) * (D(I) - D(I - Ij)$ (2.3)

It is said that $Ix \subset I$, $Ix \neq I$ is an exception set of I with respect to D and C if $SF(Ix) \geq SF(Ij)$ for all $Ij \subset I$ (2.4)

Therefore, it is defined for each $Ij \subset I$, the smoothing factor. Based on these functions, an algorithm is used which is comprised of steps to be taken in the process of analysing the sequentially input dissimilarity values scaled with cardinality function to produce a good result to test for semantic anomalous text data. The study Enhanced the Sequential exception technique to detect semantic-based text anomaly because of

the features and functionalities possessed by the technique. Some of these features are;

- Its heuristics are like random sampling or best first search is applied
- Similar idea with classical statistical approach *(k=1)* but independent from the chosen kind of distribution
- It is originally designed as a global method
- Applicable to both text and numerical data type

However, the sequential exception technique cannot be directly applied on unstructured text data. To allow SET process text, other approaches are incorporated to resolve related textual data issues such as; semantic identification in documents, the representation of text semantics, information completeness in sentences, text discourse satisfiability and the argument predicate issues faced with textual data. Before text anomaly task can be performed on documents, text needs to be well-pre-processed and represented to identify similarities or dissimilarities present in text pattern before anomalous text can be uncovered. Many issues arise as text mining techniques are employed in processing text semantics. Natural Language processing techniques is employed to carter for the pre-processing of textual data in this research before any mining task is carried out.

Universiti Utara Malaysia

2.6 Text pre-processing with Natural Language Processing (NLP)

Computational linguistics has evolved into an interesting area in the field of science and technology. Many literatures concentrated mainly on Natural Language Processing (NLP) techniques to process text semantics (Cambria & Melfi, 2015; Fyshe, Talukdar, Murphy,

& Mitchell, 2013; Kamaruddin et al., 2015; Montes-y-gómez et al., 2002; Tsai, 2007). Interesting research by Rosario & Hearst (2004), used NLP to classify semantic relations in bioscience text, it performance evaluation using confusion metrics showed total accuracy of about 74%. On the other hand, some literatures used Non-NLP for semantic text processing. Classen, Boucher, & Heymans (2011) used a text-based approach to feature modelling that provide engineers with human readable language with a rich syntax to ease modelling, but also with a formal semantic to avoid issues like ambiguity. It was concluded that language modelling approach is computationally demanding. Another research by Ray & Craven (2001) represented text structure in sentences using Hidden Markov Model to retrieve text information. It showed a significantly accurate result but was unable to capture relations in words and identifying synonyms is difficult most times.

Non-NLP techniques requires training of data and gives the burden of learning syntactic structure of any given data. Most NLP application like sentimental analysis and information extraction requires both syntactic and semantic analysis at every level. At syntactic level NLP is used to develop effective algorithm to predict part of speech tags for words sentences as well as their relationships with subject, objects and modifiers. A typical example of this is the parsing technique using algorithms like the "multilingual linear time parsing algorithm"(Hirschberg & Manning, 2015; Peter Norvig, 2015). Natural language has made use of parsing to work on small sets of discrete categories such as the Noun and Verb Phrases (NP & VP). In an attempt to capture the overall richness of linguistic expressions in documents, semantic and syntactic parser can be used for optimized identification of grammatical constructs in sentences (Berant, Chou, Frostig, & Liang, 2013).

Universiti Utara Malaysia

The goal of NLP is “to accomplish human-like language processing”. The choice of the word ‘processing’ is very deliberate and should not be replaced with ‘understanding’. For although the field of NLP was originally referred to as Natural Language Understanding (NLU) in the early days of AI, it is well agreed today that while the goal of NLP is true NLU, that goal has not yet been accomplished. There are more practical goals for NLP, many related to the application for which it is being utilized. However, NLP makes use of some of the below listed pre-processing techniques for text input. These techniques gives meaning to texts and how they are formed (Navigli, 2009a).

Pre-processing phase of text in any given context (sentences, phrases, paragraphs or a whole text document) can be represented in a more structured form of vector with different types of features that shows the relationship between words (Alagi, 2009; Kshirsagar et al. 2015; Navigli, 2009). Word representation in this context supports additional knowledge that allows the automatic identification of sense from a reference inventory. Some set of finite features are selected to represent the context in sentence. These are information resulting from the above-mentioned pre-processing steps such as part-of-speech, grammatical relations, lemmas, word parsing etc. These features are classified as follows;

- Local features are the representation of local context of word usage, it is a feature of a small number of words surrounding the target word, including part of speech tags.
- Topic features in contrast to local features defines the general topic of a discourse, thus representing more general contexts Bag of words (BOW).

Universiti Utara Malaysia

- Syntactic features are syntactic cues that has relationship with the target word and words within sentence
- Semantic features representation semantic information. A typical example is an established sense of words within a given context (Navigli, 2009).

NLP prepares text for other approaches such as Machine Learning approach to perform intended task on text data (M. J. Denny & A. Spirling, 2018). In this research semantic-based text anomaly is desired to be detected in huge numbers of document employing anomaly detection approach.

2.7 Text Similarity Measurement

This study groups semantic similarity into three categories; structure, knowledge and corpus-based metrics as thus explained;

- Structure-based metric employs a computational function of measuring the hierarchical ontology structure (is-a and has-a, part of) within a context of documents. This computational functional calculates the length of a given path connecting terms and its positional taxonomy. Thus, the closer concepts are related or similar, the more links they have among concepts (Slimani, 2013).
- Feature/ Knowledge-based similarity measures the function of term properties. A typical instance is “glosses” or definition in WordNet. This is based on the relationship or similarity relatedness of terms in hierarchical structure. There are various methods that were created to measure the degree to which words are related semantically using information drawn from semantic networks (Budanitsky & Hirst 2001).

Universiti Utara Malaysia

- Corpus-based similarity measures identifies the level or degree of similarity relatedness between words by employing information that are particularly derived from large documents or corpora. In this work, Latent semantic analysis is considered (Gomaa, 2013; Steinberger & Ježek, 2004). A comparison of text semantic similarity measures was reviewed in Table 2.3.

Table 2.3.

Text semantic similarity measures

Text Semantic Similarity with Authors	Description	Key property
Shortest path *Sim(C1,C2)=2*Max(C1,C 2)-SP* (Mihalcea, Corley & Strapparava, 2006)	This method variant is of distance-based and are particularly designed to work with hierarchies	It is hierarchically formed in terms of positioning, where by text semantics are formed based on distance, structure-based and edge counting measures. Data source Ontology
Weighted links (Mihalcea, Corley & Strapparava, 2006)	Factors which affects the weight of a link is considered in this method by computing similarity between two concepts.	Text semantics is hierarchically formed with weighted links, its data source is from Ontology and measures structure and counts edges.
Wu &Palmer (Wup) Simwup=[*(2*depth(LCS))/ (depth(concept1) +depth(concept2))]* (Mihalcea, Corley & Strapparava, 2006) (Slimani, 2013)	Two concepts similarity relatedness or measures in each taxonomy is considered. Hence, the position of both concepts is compared with the position of the most common concepts C.	Text similarity is formed semantically, it measures KB, its source is ontology-based.
Leacoock &Chodorow (Lch) (Mihalcea, Corley & Strapparava, 2006) (Slimani, 2013)	The shortest path between concepts considering the node-counting based on the shortest path between concepts.	Text similarity is formed semantically KB Data source is ontology based.

Resnik (Res) (Slimani, 2013)	Information concepts of shared parents are considered in this method	KB Data source is ontology based and corpus. Text similarity is formed semantically
Lin (Slimani, 2013)	Lin measures is developed on Resnik's measure of similarity, thereby adding a normalization factor that consists of the information content of the two inputted concepts	KB Data source is ontology based and corpus. Text similarity is formed semantically
LSA (Steinberger & Ježek, 2004) (Turney & Pantel, 2010) (Slimani, 2013)	Co-occurred terms found in corpus are captured using dimensionality reduction approach (SVD) on a term-by-document matrix T representing corpus.	Corpus-based Data source. Text similarity is formed semantically
PMI-IR Point wise mutual information using data collected by information retrieval (Mihalcea et al., 2006) (Turney 2001)	It evaluates the semantic similarity of words using an unsupervised measure that is based on counting word co-occurrence of large corpora.	Corpus-based Data source. Text similarity is formed semantically
Lesk (Slimani, 2013)(Sardar, 2018)	These methods allow the definition of two similar concepts as a function of the overlap between the corresponding definitions. However, it can be employed with dictionary definitions and it is not limited to semantic networks.	KB Data source is ontology based and corpus. Text similarity is formed semantically
Pairwise word measures (Mihalcea, Corley & Strapparava, 2006)	It operates on graph. It only considers best-match average of similar words.	Corpus-based Data source. Text similarity is formed semantically

Universiti Utara Malaysia

After reviewing the advantages and disadvantages of each similarity measures in Table 2.3, it is imperative that knowledge-based approach offers high precision but has difficulties in dealing with overlap sparsity and it is dependent on resources like the dictionary. While the corpus-based approach with respect to implementation is better than other approaches. However, the results of corpus-based approach are not satisfactory with resource scarce language. It is best to integrate both knowledge and corpus-based approach for the study objective.

A similarity measure is a function that computes the degree of similarity between pairs of text object (Gomaa, 2013). There are a large number of similarity measures proposed in many literatures to measure similarity in text documents (Cha, 2007; Huang, 2008; Mihalcea et al., 2006; Slimani, 2013; Yih & Meek, n.d.). In agreement with the above literature about similarity in text, text clustering is one of the most effective ways of measuring text similarity. Its accuracy requires a precise definition of the closeness between pairs of objects, in terms of either pair wised similarity or distance (Huang, 2008; Mihalcea et al., 2006).

Universiti Utara Malaysia

A variety of similarity measure have been proposed and widely applied, such as cosine similarity, overlap coefficient and jaccard similarity (Cha, 2007; Yih & Meek, 2009). All similarity measures map to range (-1, 1) or (0, 1) shows minimum similarity (incompatible similarity) (Huang, 2008). One (1) shows absolute or maximum similarity while zero (0) or negative one (-1) shows absolute dissimilarity.

Some measures which have been popularly adopted for computing the similarity between two documents are briefly presented here. Let *d1* and *d2* be two documents represented as

vectors. The Euclidean distance measure is defined as the root of square differences between the respective coordinates of d_1 and d_2, (Cha, 2007; Lin, Jiang, & Lee, 2014).

$$d_{Euc}\,(d_1, d_2) = [(d_1 - d_2) \cdot (d_1 - d_2)]^{1/2} \quad (2.5)$$

where $A \cdot B$ denotes the inner product of the two vectors A and B. Cosine similarity measures the cosine of the angle between d_1 and d_2 as follows:

$$S_{Cos}\,((d_1, d_2) = (d_1 \,.\, d_2) / [(d_1 \,.\, d_1)^{1/2}(d_2 \,.\, d_2)^{1/2}] \quad (2.6)$$

Pairwise-adaptive similarity dynamically selects a few features out of d_1 and d_2 and is defined to be

$$d_{Pair}\,(d_1, d_2) = (d_{1,K} \cdot d_{2,K}) / [(d_{1,K} \cdot d_{1,K})^{1/2}(d_{2,K} \cdot d_{2,K})^{1/2}] \quad (2.7)$$

The Extended Jaccard coefficient is an extended version of the Jaccard coefficient for data processing (Cha, 2007; McInnes & Pedersen, 2013; Szmeja et al., 2018):

$$S_{EJ}\,(d_1, d_2) = (d_1 \,.\, d_2) / [(d_1 \,.\, d_1 + d_2 \,.\, d_2 - d_1 \,.\, d_2)] \quad (2.8)$$

Universiti Utara Malaysia

while the Dice coefficient looks like it and is defined as follows:

$$S_{Dic}\,(d_1, d_2) = (2d_1 \,.\, d_2) / [(d_1 \,.\, d_1 + d_2 \,.\, d_2)] \quad (2.9)$$

IT-Sim, an information-theoretic measure for document similarity as seen below,

$$S_{IT}\,(d_1, d_2) = (2\textstyle\sum_{wi} min\,(p_{1i}, p_{2i})\, log_{\,\pi(wi)}) / (\sum_{wi} p_{1i}\, log_{\,\pi(wi)} + \sum_{wi} p_{2i}\, log_{\,\pi(wi)}) \quad (2.10)$$

where wi represents feature $_{i,}$ p_{ji} indicates the normalized value of w_i in document d_j for $j = 1$ or $j = 2$, and $\pi(w_i)$ is the proportion of documents in which w_i occurs (Lin et al., 2014).

i. Manhattan Distance also known as the city block distance. This similarity measure is the distance between two points. it is the sum of the absolute differences of their Cartesian

co-ordinates. In a simple way of saying it is the total sum of the difference between the *X*- co-ordinates and *Y*- co-ordinates. Suppose two points *A* and *B* to find the Manhattan distance between them, to sum up the absolute *X*-axis and *Y*- axis variation means to find how these two points *A* and *B* are varying in *X*-axis and *Y*-axis. In a more mathematical way of saying Manhattan distance between two points measured along axes at right angles. In a plane with *P*1 at *(x1, y1)* and *P*2 at *(x2, y2).*

$$\text{Manhattan distance} = |x1 - x2| + |y1 - y2| \tag{2.11}$$

Manhattan distance metric is also known as Manhattan length, rectilinear distance, *L*1 distance or *L*1 norm, city block, distance (Gomaa, 2013) .

ii. Eugene Euclidean distance is best at measuring textual data when it is dense or continuous. The Eugene Euclidean distance between two points is the length of the path connecting them. The Pythagorean theorem gives this distance between two points (Karkali et al., 2014). It satisfies all the above four conditions and therefore is a true metric. It is also the default distance measure used with the K-means algorithm. Measuring distance between text documents, given two documents *da* and *db* represented by their term vectors ta and *tb* respectively, the Euclidean distance of the two documents is defined as;

$$\text{DE (ta, tb)} = \left(\sum_{t=1}^{m} (\omega_{ta} - \omega_{tb} 1^2)^{1/2} \right) \tag{2.12}$$

Where the term set is *T = {t1, . . . , tm}*. As mentioned previously, the use of *tfidf* value as term weights, i.e. *wt ,a = tfidf (da , t)*

Universiti Utara Malaysia

iii. Jaccard Dissimilarity is based mainly on sets. It is also known as the Jaccard coefficient which is a similarity measure that ranges between 0 and 1.

$$d_J(A,B) = 1 - J(AB) = \frac{|A \cup B| - |A \cap B|}{|AUB|} \quad (2.13)$$

- Sets are an unordered collection of objects *{a, b, c} = {c, b, a}*.
- Cardinality of *A* denoted by $|A|$ which counts the elements in *A*.
- Intersection between two sets *A* and *B* is denoted $A \cap B$ and reveals all items which are in both sets *A, B.*
- Union between two sets *A* and *B* is denoted $A \cup B$ and reveals all items which are in either set (Gomaa, 2013; L. I. N. Li, Hu, Hu, Wang, & Zhou, 2009).

The study aims to compare the results of the selected measures to know which measure benefits the ESET best.

Universiti Utara Malaysia

iv. Cosine dissimilarity measures identifies the normalized dot product of two attributes. By determining the cosine similarity, to effectively find the cosine of the angle between two objects. The cosine of 0^o is *1* and it is less than *1* for any other angle. It is thus a judgement of orientation and not magnitude: two vectors with the same orientation have cosine similarity of *1,* two vectors at 90^o have dissimilarity of *0,* and two vectors diametrically opposed have dissimilarity of *-1,* independent of their magnitude. Cosine dissimilarity is particularly used in positive space, where the outcome is neatly bounded in *[0, 1]* (Deshpande et al., 2014; Gomaa, 2013; Henriksson et al., 2014; Huang, 2008). One of the reasons for the popularity of cosine dissimilarity is that it is very efficient to evaluate, especially for sparse vectors.

$$sim\ (A, B) = cos\ (\Theta) = |\ A.\ B| / ||\ A||\ ||B|| \quad (2.14)$$

In this study, cosine similarity is considered because, it measures similarity of features in Vector Space Model (VSM). This helps in generalizing the idea that cosine similarity is a metric that measures the orientation and not the magnitude of words. It is a comparison between documents on a normalized space. For instance, the application of cosine similarity in the field of NLP is quite intuitive. Word features such as n-gram can be similar though formally they are considered as different features. The word *game* and *play* are different words with different VSM points yet semantically related. In cases like this, WordNet can be applied to multiply VSM matrix to compare similarity of words with different mapping points (Lin et al., 2014). A multitude of measures for computing similarity between text semantics have been proposed over the years and many ongoing research are in process(Gomaa, 2013; L. I. N. Li et al., 2009; McInnes & Pedersen, 2013; Szmeja et al., 2018). These words were then compared with an external corpus or dictionary to evaluate performance.

Universiti Utara Malaysia

This work contributes to the research by proposing an enhanced LSA similarity index for identifying semantic anomalous textual data as a Corpus based metric. It compares text document using SVD. LSA have been used in many research to uncover semantics in text document (Froud et al., 2013; Steinberger & Ježek, 2004; Yan, Guo, Lan, & Cheng, 2013). Furthermore, it was noticed from previous research that LSA can be employed to measure the effectiveness in producing more coherent clusters in text documents (Froud, Lachkar & Ouatik.2013).

2.8 Text Semantics

The meanings assigned to any stretch of text, phrase, sentence or documents are of course a function of many contexts. The analysis of those contexts and how meanings are assigned to them differently is the semantics in general. According to (Chandola, Banerjee, & Kumar, 2009), detection of semantic in an unstructured text data tells us the reasons why meaning representation is needed and how it can be of help in knowledge acquisition and decision making. In the process of trying to understand text from corpus cannot be directly linked with the events in the text into a sequential structure. Classically in Artificial Intelligence, inference be filling in the missing connections between the surface structures fragments of text by recourse to context and knowledge about a given document. Semantic Analysis is an approach in text mining that helps in representing knowledge using language. Language is a very generic representation, where words are used to describe almost anything (Beltagy et al., 2015; Sun, Guo, Lan, Xu, & Cheng, 2016). Usually, semantically related information can be obtained manually from sentence on three aspects, these aspects includes; the objects the sentence describes, the properties of these objects and the behaviours of these objects. To measure sentence similarities from those three aspects, Objects-Specified Similarity is defined to demonstrate the similarity between the objects which sentences describe; Objects-Property Similarity shows the similarity between objects properties of sentences while Objects-behaviour similarity expresses the relationship between objects behaviours. Overall Similarity denotes the overall relatedness of sentences, which combines the above three. Similarity calculations involves sentence chunking and semantic similarities between words. Chunker divides sentence into series of words that compose of a grammatical unit (mostly

Universiti Utara Malaysia

noun, verb, or preposition phrase). While semantic similarity between words form vectors and computing their similarities is as follows.

- All nouns (objects specified) derived from noun phrases of a sentence into an object specified vector is conceptually like a typical vector space representation used in a standard Information Retrieval approach. Each entry in the vector is derived from computing the word similarity between a word feature w_i and each noun from noun phrases of a sentence.
- After that, the maximum score from the matching words that exceeds certain similarity threshold θ will be chosen. Here we take $\theta = 0.2$.
- Suppose sentence s_1 and s_2 are the two sentences to be compared and there are two noun phrases in s_1, which are $np_{11}= \{w_1, w_2, w_3, w_4\}$ and $np_{12}= \{w_1, w_5, w_6, w_4\}$, and the nouns in np_{11} and np_{12} are $noun_1= \{w_3, w_4, w_6\}$. There is one noun phrase in s_2, which is $np_{21}= \{w_1, w_7, w_3\}$, the nouns in np_{21} are $noun_2= \{w_3, w_7\}$.
- Words in $noun_1$ forms 1's objects specified set. Words in $noun_2$ forms 2's objects specified set. The feature set of objects specified vector is $vf_{os}= noun_1 \cup noun_2 = \{w_3, w_4, w_6, w_7\}$. For each word w_i in the vector entries, the formations of objects specified vector v_{os1} and v_{os2} of s_1 and s_2.

The heterogeneity and inconsistency of text data makes it difficult to detect semantics when trying to represent knowledge. These are mainly concerned with morphological issues in text. Practical semantic analysis systems adapt strictly to compositional approach by relying on semantic grammars and extracting information from source. This approach is based on the principle of compositionality "meaning of a sentence can be composed from meaning of its parts" i.e. Part-of-Speech of a sentence (Jurafsky & Martin, 2000;

Lenci, Montemagni, & Pirrelli, 2001; Gahl et al. 2003) . Many approaches or models can be performed to analyse, construct concepts, ideas or topics with identified text semantics in documents. As will be discussed in the next sub-section.

a. Word / document matrix simply means word and document having relatively close meaning or semantics. Semantic vectors have to main functions namely; Indexes needs to be created to build word space models and Searching through vectors in models. One way to simply explain SVD to an audience with little or no idea about linear algebra is that the SVD of a matrix A (which may be non-square) comes from diagonalizing the square, Hermitian matrix $\begin{bmatrix} 0 & A \\ A* & 0 \end{bmatrix}$ Where 0 denotes a zero matrix of the appropriate size. Then the eigenvectors of this extended matrix are given by the pair of

   $(u_i, \bar{v}_i)$ $(ui, \bar{v}i)$, *with* eigenvalues $(\sigma_I \sigma_{i,})$ so $(u_i u_i)$ and $(v_i v_i)$ are given, this form $(v_i v_i)$ $SVD\ A = U \Sigma V * A = U \Sigma V *$ Singular Value Decomposition (*SVD*)

   Eqn2. 1 (2.15)

   SVD has two wonderful properties that make it very helpful and important. First it is best known for its existence for *any* and *all* matrices: large, small, square, rectangular, singular, non-singular, sparse and dense. The next helpful property is the optimality property. This property enables the ranking approximation of matrix. To get the singular value decomposition, is an advantage of the fact that for any matrix

   A, $A^T A$ is symmetric since $(A^T A)^T = A^T (A^T)^T = A^T A (A^T A)^T = A^T (A^T)^T = A^T A)$ (2.16)

b. LDA model improves mixture models that capture the exchangeability of both words and documents from the old way by PLSA and LSA (Gomaa, 2013). the basic idea of

the process is, each document is modelled as a mixture of topics, and each topic is a discrete probability distribution that defines how likely each word is to appear in each topic. These topic probabilities provide a concise representation of a document (Blei, Ng, & Jordan, 2012; Z. Liu, 2013).

c. Non-negative Matrix Factorization (NMF) is a low-rank approximation technique which introduces the constraint that the data matrix and the factorizing matrices are nonnegative. By allowing only additive linear combinations of components with nonnegative coefficients (i.e. a conical combination), NMF inherently leads to a parts-based representation.

Compared to PCA, which is a holistic representation model, NMF leads to a much more intuitive and interpretable representation. Formally, NMF is characterized by the following factorization; In this process, a document-term matrix is constructed with the weights of various terms (typically weighted word frequency information) from a set of documents. This matrix is factored into a term-feature and a feature-document matrix. The features are derived from the contents of the documents, and the feature-document matrix describes data clusters of related documents. it automatically clusters the columns of input data. It is this property that drives most applications of NMF.

Foltz (1996) made use of Latent Semantic Analysis (LSA) to compare semantics similarities in text data. The experiment showed that comparison made by LSA are like approaches that use propositions in the sense that they both make comparisons at semantic level rather than surface feature level. Another research performed by Brants

et al.(2002) made use of the probabilistic latent semantic analysis to segment document based on their topics.

This research combines the use of (PLSA) model with a selecting segmentation point model based on the similarity value between pairs of adjacent blocks. Results on commonly available data are significantly better than those of other State-of-the-art system. Leveling (2007) stated that, semantic analysis resolve co-reference, decompound words in sentences and identifies metonymy, synonym and polysemous words. Hassan & Mihalcea (2011) made research on how identified meaning of words can be characterized by salient concepts found in its immediate context. The new built model in this research was found to be computationally efficient.

Topic models can give an understanding into the hidden latent structure of an extensive corpus of documents. A scope of methods has been proposed in the writing, including probabilistic topic models and techniques based on matrix factorization. In Belford, Mac Namee, and Greene, (2018) demonstrate the inherent instability of popular topic displaying approaches, utilizing a few new measures to evaluate dependability. To address this issue with regards to matrix factorization for topic demonstrating, steadiness of topic displaying through matrix factorization was utilized on ensemble learning techniques. Based on investigations performed commented on content corpora, demonstrated that a K-Fold ensemble technique, consolidating the two ensembles and organized introduction, can significantly lessen instability, while at the same time yielding more precise topic models.

d. According to Froud et al. (2013) LSA gives a superior proportion of coherent clusters in content documents. LSA looks at content archive gathering and contrasts it and

other comparative documents. Thusly, it thinks about related documents to be semantically related and considers documents with couple of regular words as semantically distant. Chang, Yih, and Meek, (2013) demonstrated a methodology that coordinates Multi Relation Latent Semantic Analysis (MRLSA) from homogenous and heterogeneous data sources.

This methodology accomplished a best in class execution on other existing benchmark data for two relations. This straightforward strategy associates shockingly well with how a person, taking a gander at substance, may classify record accumulation. This exploration thinks about the comparability of sentences to upgrade the revelation of irregular literary information in corpus. The unavoidable issue is how well does LSA functions? This inquiry can be replied in two unique folds.

LSA fills in as a portrayal show and as a similitude measure for human verbal ideas (Landauer, Foltz, and Laham, 1998). All in all, LSA is connected on tremendous accumulation of corpuses (in excess of 10000 documents) based on the recognizable proof of relationship between words. On the off chance that content archive gathering is pretty much nothing, no meaningful relationship between words can be inferred. Another more controlled medium utilizing LSA is to perform archive development. It takes related words that are gotten from record and to specifically extend the first archive with more words from LSA. This may not work similarly well for all documents or all words. Documents whose topics command in the accumulation will probably be enhanced, while strange record or word will be mapped to as an abnormality. Uncommon words won't have meaningful comparable words regardless, either LSA is utilized or not. Even though, LSA has been utilized generally with fruitful experimental outcomes, its downside is in the low

Universiti Utara Malaysia

rank guess where it needs elucidation and the outcome will be less compelling for vast heterogeneous content accumulations (Kamaruddin, 2011). Note that comparability gauges inferred by LSA are not basic contiguity frequencies, co-event include, or relationships utilization, it relies upon a great mathematical examination known as Single Value Decomposition (SVD) (Chang et al. 2013; Foltz, 1996; Steinberger and Ježek, 2004).

SVD is able to do effectively deducing considerably more profound relations (in this way the expression "Latent Semantic"), and as outcome are often much better indicators of human importance based on judgments and execution than are the surface level possibilities that have for quite some time been dismissed by etymologists as the premise of dialect wonders (Lund and Burgess, 1996). This examination utilized LSA on account of its promising outcomes in recognizing semantics in huge documents. Be that as it may, to determine issues identifying with content uncertainty, content Canonical Form was utilized utilizing WSD calculations. This will be clarified in the following area.

Universiti Utara Malaysia

2.9 Text Canonical Form

Canonical Form is a notion that greatly simplifies task by dealing with a single meaning representation for an extensive variety of expression (Jurafsky and Martin, 2000). Mittal, Moorthy, and Bovik, (2012) stated that there are three factors that characterizes Canonical Form in sentences which are words syntactic structures, semantic role and the recurrence of usage. More in this way, Bates et al., (1988) noticed that sentences with Agent-Action Object arrange speak to the canonical word for grammar.

Distinctive approaches have been utilized in existing literatures, for example, synonyms

(Kamaruddin et al. 2012) and segmentation ambiguity (Abouzakhar et al., 2008; Montes-y-gómez, Gelbukh, and López-lópez, 2002) in this way, these approaches are noticed to be costly and computationally complex when utilized. CF lay more emphasis on how knowledge that can be displayed in a short manner to correctly recognize faculties in expression. Frameworks like the GRAIL, SNOMED and GALLEN are planned by utilizing CF representation procedure to express knowledge meaning without the appeal to a special intervention of knowledge interpretation in medical informatics (Baud et al. 1993; Galeazzi, Mori, Consorti, Errera, and Merialdo, 1997). Galeazzi et al., (1997) structured a philosophy to assign responsibilities to regulate interactions safeguarding cognizance that is essential to appropriate basic procedure of demonstrating an individual concept, from the expression in the original corpus to the CF in the GRAIL display among cooperating focuses. The system was applied effectively to coordinate the demonstrating activities of three other teams of surgeons in Rome. Spackman, (2001) proposed in his research that, normal forms can be utilized to describe rationale expressions of clinical concepts in SNOMED RT. The paper describes variations between decision of syntax with normal form and definitions of several normal forms including short and long canonical forms. It was concluded that definition in short CF can be utilized in conveying XML-syntax definitions of concepts in SNOMED RT while definitions in Long CF cannot be inputted in a relational table with one segment for each role and one line for each concept. The research expects to streamline communication about description rationale definitions for clinical wording.

CF shows a systematic relationship between word faculties and the grammatical constructs found in content documents. On account of content variants sharing similar

name element, for example, Mr James and James Bond because of the standard English naming convention (Cresswell, 2016). Each name is categorized as an element compose with a canonical name as its identifier. The canonical name is the fullest least ambiguous label that can be alluded to in content archive.

CF was employed in this work by utilizing WSD and two of its algorithms namely; Lesk and Selectional inclination leveraging their potential expressly in the identification of content semantics in documents. In this manner, will perform other functions like;

- Reducing inputted text data with multiple meaning into a single common word.
- Equating words in sentence implicitly to agents, action and object, rather than subject, verb and object.

Language in general is ambiguous, so that words can be said, interpreted and understood in different ways depending on its context of occurrence. Unfortunately, identifying a specific meaning that a word, sentence or phrase assumes in a context seems easy. While many people do not even know or think about the ambiguities of language, unstructured textual data needs to be well processed and transformed into structures which must be analysed to determine the underlying meaning of information within a context.

It is described as an Artificial Intelligence (AI) complete problem by analogy to Natural Language Processing (NLP) completeness in complexity theory, a problem whose solution solves central problems of Artificial Intelligence (AI). Research by (Tan, Zhang, Clarke, & Smucker, 2015), used Lexical similarity measures to resolve the coordination of ambiguity in documents. Two similarity system was proposed in the research, which are WordNET and the second was the distributional extracted information using the C&C

Universiti Utara Malaysia

parser to extract information from Wikipedia. Similarity of syntactic structures and head words was considered. Since similarity is appropriate for handling the coordination of conjuncts which are likely to be alike semantically. Although this is not a general rule because lexical items sometimes may not be semantically similar (Eshghi, Howes, Gregoromichelaki, Hough, & Purver, 2015). More so, Text annotation aligned with bilingual text and monolingual text using statistics of lexical association was experimented to coordinate ambiguity in text. However this method operates without semantic annotation and works only at the text level (G. Zhou et al., 2011). Word Sense Disambiguation (WSD) basically relies on knowledge in fact, the fundamental procedure of any WSD system is as follows;

- BOW or sentence from a context that makes use of knowledge from different sources
- Knowledge source vary from different collection of texts, these texts collection can be either unlabelled or labelled word senses from a more structured resource like semantic network. It is impossible for human and machine to identify useful meaning within a given context of information without knowledge. Unfortunately, manual creation of knowledge resources involves expensive and time-consuming effort. This problem is known as the knowledge acquisition bottleneck (Navigli, 2009).

WSD is an automatic classification technique, where Word senses are viewed as the classes, these classes are assigned to every occurrence of one or more classes based on the evidence from resources either external knowledge or context sources. Other studies of classification tasks can be seen in the field of NLP such as part of speech tagging, named

entity resolution and text categorization (Jurafsky & Martin, 2000; Navigli, 2009a). the difference between NLP and WSD task is that NLP uses the above-mentioned task (part of speech tagging, named entity resolution and text categorization) while WSD depends on a finite set of classes that changes depending on word to be classified. Therefore, it is said that WSD comprises of *n* distinct classification tasks, where *n* represents the total size of lexicon in each context. Variants of the generic WSD task can be divide into two; targeted WSD and the All-words WSD. Tests sets, knowledge sources are expected to have different features in terms of their performance and operations. Table 2.4 gives a summarized explanation of the three approaches (Navigli, 2009).

Table 2.3.

Word sense disambiguation approaches

Approach with algorithms	Opportunities	Challenges
Knowledge-Based Lesk Algorithm, Semantic Similarity, Selectional preference Heuristic	High precision is achieved using this approach.	These algorithms have difficulty in dealing with overlap sparsity, since it is Overlap-based, and its performance depends on resources like Dictionary definitions.
Supervised Decision list, Decision tree, Naïve Bayes, Neural Networks, Exemplar-based or Instance-based learning, Probability mixture, Rank-based combination, SVM, Ensemble method, Majority voting AdaBoost	With respect to implementation perspective, this approach is better than others.	Its result on resource scarce languages are not satisfactory compared to other approaches.
Unsupervised Context clustering, Word clustering, Co-occurrence graph Spanning- tree based approach	In this approach there is no need of sense inventory or sense annotated corpora.	Result performance is always not as good as other approaches, and it is difficult to implement

Source (Navigli, 2009)

As stated earlier, two WSD algorithms were combined from the supervised approach

namely; Lesk and Selectional preference algorithm. Reasons for choosing these WSD algorithms will be further explained as the study continues.

a. LESK ALGORITHM

Lesk algorithm uses dictionary definitions (gloss) to disambiguate a polysemous word in a sentence context (Lesk, 1986). The major objective of this idea is to count the number of words that are shared between two glosses. The more overlapping the words, the more related the senses are. To disambiguate a word, the gloss of each of its senses is compared to the glosses of every other word in a phrase. A word is assigned to the sense whose gloss shares the largest number of words in common with the glosses of the other words (Banerjee, 2002; Torres & Gelbukh, 2009). Due to the high dimensionality of search spaces, Gelbukh, Sidorov, & Han, (2005) made use of some heuristics to find a near optimal combination of intended senses present in text documents. Some lesk-like WSD algorithms are considered as a global coherence average relatedness between chosen words in documents. Another research by Torres & Gelbukh, (2009) compared several similarity measures applied to lesk WSD algorithm. However, the research showed that similarity relatedness of words in text documents performs better with lesk. Basile, Caputo, & Semeraro, (2014) enhanced lesk WSD algorithm through distributional semantic model by extending two well-known variations of lesk WSD methods namely words and contextual information from source documents.

Lesk algorithm exploits the concepts of maximum overlap between the context of a word and its definition of its senses to select proper meaning. The evaluation performed on SemEval-2013 Multilingual WSD shows that the enhanced lesk algorithm goes beyond

Universiti Utara Malaysia

the most frequent sense baseline and the simplified version of the Lesk algorithm. Moreover, when compared with the other participants in SemEval-2013 task, the proposed approach could outperform the best system for English.

b. SELECTIONAL PREFERENCE ALGORITHM

Selectional preference algorithms tries to capture the linguistic arguments elements of a certain semantic class in a large corpus. A corpus-based approach for selectional preference extracts (verbs/subjects) relations from large corpus and as well use an algorithm to generalize nouns from each verbs separately (Agirre & Martinez, 2002). Wagner, (2000), introduced an automatic modification of the selectional preference approach for ontology building to represent lexical semantic knowledge by means of statistical corpus analysis.

Selectional preference algorithm is an alternate method of producing abstract concepts of verb argument (Gong, Zhao, & Zhu, 2016). Another approach or method was presented for examining the abilities Neural Machine Translation (NMT) with Word Sense disambiguation. It was however, noticed that there were still much work to be done to improve WSD on NMT (Marvin, 2018). An interesting work performed by Sayeed, Greenberg, & Demberg, (2016) leveraged the thematic fit evaluation for selectional preference to find the correlation of computer outputs judgment with human- collected judgement in the context of real valued vector space word representation.

This approach is applicable when predicate-argument relations are used to perform important task. However, word semantics disambiguation in corpus is incomplete if the core semantic is not well represented. Another study made by Chaplot & Salakhutdinov,

(2018) employed the knowledge-based WSD approach on topic model which incorporates semantic information about synsets as its priors. The proposed model scales linearly with the number of words in the context, which allows the use of document as the context for disambiguation and outperform state-of-the-art knowledge based WSD system on a set of benchmark data. This study combined both Knowledge-based and Corpus based word disambiguation approach to disambiguate senses from text documents.

2.10 Semantic Representation Scheme

Over the years, NLP techniques have been able to detect text structure samples. More so, computational identification of relationships between words still needs to be investigated upon since its problem is much harder (Jurafsky & Martin, 2000; Parr, 2012). Text representation is the most common and basic method of transforming information into linguistic structure, it aims at generating a formal representation of identified sentences in order to uncover semantics (S. S. Kamaruddin et al., 2015). Discussion continues in this study to review other text representation schemes such as the language modelling, graph-based representation and others.

a. Vector Space Model and Term Frequency

M. Liu and Yang (2012) recommended an upgrade of TF-IDF weighting in the vector space model for improvement of classification accuracy by introducing a new parameter to represent the in-class characteristic; the new weighting method is called TF-IDF-CF based on TF-IDF. Traditional TF-IDF employs a term weighting scheme that is based on term ontology (Punitha, 2012). The reason for this use is that TF-IDF only pays attention to the repeated words in the document and disregards the other factors that may influence

the world weighs. Huang and Wu, (2013) proposed an improved TF-IDF algorithm to solve the low accuracy of micro-blog commercial word extraction and application it in term weight calculation. The possibility of applying the improved algorithm involved classifying a large amount of micro blog information into certain patterns and then assigning term weights for the classes under the Hadoop distributed framework by using the improved TF-IDF algorithm. The results indicate that the application of the improved TF-IDF algorithm in micro-blog commercial word extraction is effective and enforceable.

b. Graph Based Representation

Chakravarthy, Venkatachalam, and Telang (2010) proposed a graph-based approach that employs two domains -independent graph representations to cover text (webpages and email). The graph representations are selected based on domain knowledge to provide focus on the domains. In the case of similarity between documents in different clusters, the edge appears between two nodes. If the documents that are contained in the cluster are highly related, the edges in the same cluster will weigh more than the edges across clusters. Each graph-based algorithm may produce the ground differently. They may also use graph partitioning algorithms differently (Kaur & Kaur, 2013).

c. Conceptual Graph Representation

Conceptual graph is a knowledge representation language. It is based on the field of linguistics, psychology and philosophy (Sowa & Way 1986). Conceptual graphs (CGs) are used to represent knowledge structures at semantic level. CGs are finite, connected, bipartite (Involving two elements: concepts and relations) graphs. A graph is comprised of a set of vertices or nodes and edges. Diagrammatically, it is depicted as a collection of

nodes and arcs.

The concept nodes represent concepts such as entities, attributes, states and events while the relation nodes represent relations to show how the concepts are interrelated. The arcs are used to link the concept nodes to the relation nodes. As a contrary to other network languages, the edges are not labelled. These edges can be weighted to show its importance and to facilitate further manipulation of the graphs but in this research simple graphs are favoured because it is enough to efficiently represent the problem domain. An example of CG to represent the sentence "The dog is sitting on a couch" is shown in Figure 2.3. As shown in Figure 2.3, the concept nodes are drawn as a box and the relation nodes are drawn as a circle. The arcs are drawn as an arrow that links the box to the circle.

Figure 2.3. Conceptual Graph Representation

2.10.1 Other representation scheme

Poon & Domingos (2010) propose Ontological USP as a system that learns an IS-A hierarchy over clusters of logical expressions and populates it by translating sentences to logical form. Ontology is text-based representation that is good for textual data

visualization but still have some limitations like; it needs for distance measure to compare ontology, it has the inability to link different areas of specialty within a given area or discipline and it ignores some words during the construction of ontology.

The experimental results obtained by Wang, Ni, Sun, Tong, & Chen (2011) showed that the dependency graph algorithm exhibits the best performance in a specified document clustering among methods that are based on the BOW model. This algorithm identifies causal relationships and improves the performance of the similarity measure between texts. Its major issue is that comparing graphs can be computationally complex (H. Liu et al. 2013; Montes-y-gómez et al., 2002).

Ma et al. (2012) developed a text mining approach based on ontology to collect novel research proposals based on the similarity in the areas of research.

The approached proved to be effective and efficient in compiling research proposals with English and Chinese texts. An ontology study was conducted for the classification of the concepts of disciplines in terms of the various regions and forming relations among them. Text mining and optimized technologies were adopted to compile research proposals based on similarities; eventually, balance was achieved in accordance with the characteristics of applicants. Experimental results showed that the proposed method improves the similarity of the proposal sets and promotes efficient of assembly in the proposal process.

2.10.2 First Order Logic (FOL)

Logic representation contains individual variables and quantifiers in analysing natural language. It is of two parts, which are; the syntax which is a group of legal construct or

expression and the semantics which refers to the meaning of these expressions. Semantics in predicate logic checks the validity of statement using the truth table. Value is attached to variable in an appropriate context to a truth table through predicates *P (t 1, t2, . . . , tn)* for precise and consistent interpretation of words (Kobus, Yvon, & Damnati, 2008).

According to Jurafsky & Martin, (2000) FOL is a well-defined computational and understandable approach to the representation of knowledge that satisfies the rules of grammatical representation in language. It is also an interpretation of grammatical construct in a language that is assigned to all constants (Non-logical) in that language (Wehmeier, 2004). An interesting research introduced the use of probabilistic logic with expressivity and automated inference provided by logical representation to capture semantics in sentences. This research demonstrated a state-of the-art performance in identifying semantics (Beltagy et al., 2015).

Garrette, Erk, & Mooney, (2014) defines the link between vector spaces through lexical mapping of predicate symbols (lexical semantics) with predicate logical forms. The resulting approach could solve many difficult textual entailment problems that requires handling complex combinations of semantic phenomena.

To resolve the problem of text identification performance, solution often depends on text representation. However, text representation schemes find it difficult to uncover useful information due to the sparsity and high dimensionality present in textual data (Chen & Wu, 2012; S. S. Kamaruddin et al., 2012; R. Kannan et al., 2017) Researchers have developed different approaches by optimizing text representation to overcome sparsity and high dimensionality in text (S. S. Kamaruddin et al., 2015, 2007, 2012; H. Liu et al.,

2013; Montes-y-gómez et al., 2002a). FOL is used in this study with concept network graph to represent semantic-based text anomaly because of the simple approach they use in identifying idea in large documents. This approach is easy to perform and gives a clear picture of the main idea that is needed to be represented.

2.11 Research Gap

The proposed research study is aimed at Enhancing Sequential Exception Technique (ESET) which was originally developed as SET. SET detects anomalies in large log files from database (Arning & Rakesh, 1996; Z. Zhang & Feng, 2009). The enhancement of SET into ESET includes several tasks which is solely aimed at detecting semantic-based text anomalous data in large numbers of document efficiently for knowledge sharing and decision-making purposes. As can be seen in Figure 2.4.

Universiti Utara Malaysia

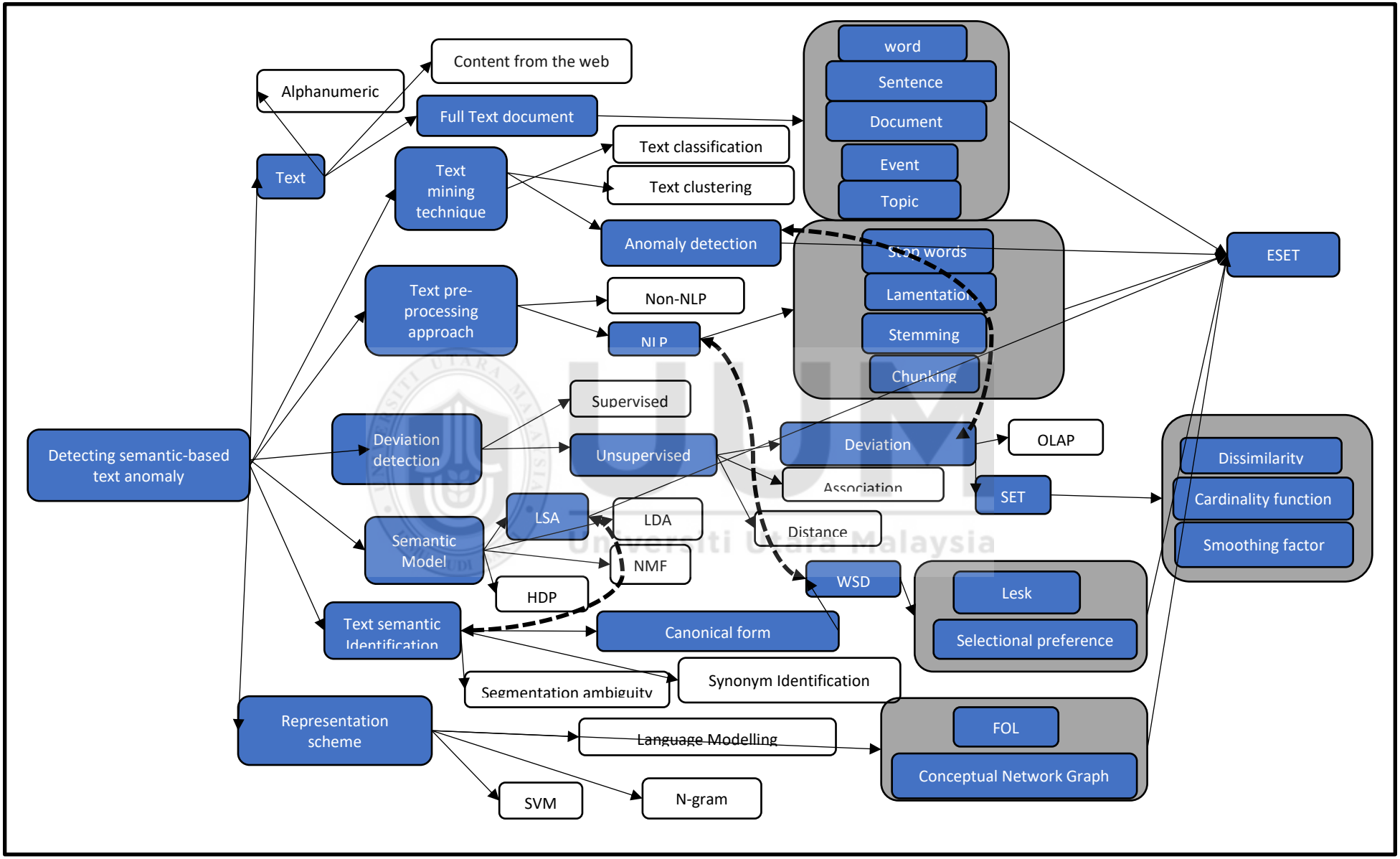

Figure 2.4. Mind-map of the study technique ESET

2.12 Summary

Chapter Two reviews literatures that are relevant to the study like: Information overload, Text Mining, Canonical form with WSD, Text semantics representation, semantics analysis in text and anomaly detection. Also, from the reviews, these literatures formulate the research framework of the study. The chapter details out text mining techniques, the essence of text processing and how anomaly-based semantics can be identified in textual documents. This study employs a novel technique by combining several approaches that will be further explained in the research methodology. However, this chapter has provided a research map to which ESET is designed and its functionalities. The literature review serves as a systematic study of what ESET entails and how it works. Notwithstanding, literatures from existing studies may have different perspective or motifs on providing solutions to text mining problems. It is essential to put ESET to test by knowing the best method of research that befits the detection of semantic based text anomalous data in documents using ESET.

Universiti Utara Malaysia

CHAPTER THREE

METHODOLOGY

3.1 Introduction

As regards to the existing studies reviewed in chapter two, related problems were defined and critical analysis relating to the outlined problems were reviewed. The summary of literatures was also shown in tabular form to identify research gaps and to determine feasible possible contributions. This has enabled the formulation of research design, experimental analysis and evaluation techniques that was used in this study.

Universiti Utara Malaysia

3.2 Research Design

ESET adopts an experimental research design which is the quantitative approach used in conducting basically the research experiment that established the cause and effect of detecting semantic based text anomalous data from huge numbers of documents. This study involves phases that gives an elaborate insight to ESET as shown in Figure 3.1.

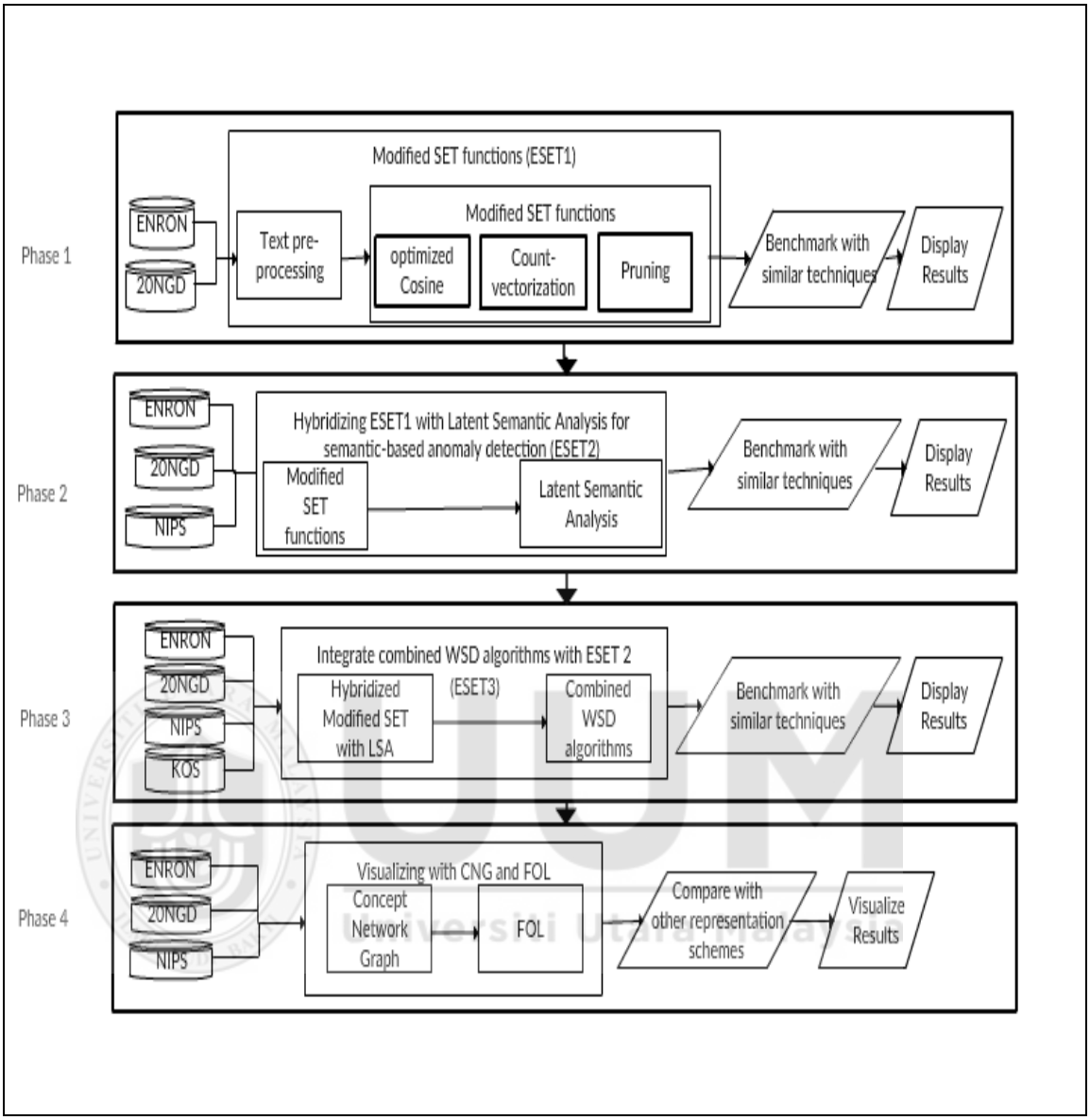

Figure 3.1. Research design for ESET

Figure 3.1. is an illustration that shows the study design model and how it was divided into phases. These phases explain the performance and phases of the technique according to the research objectives. A detailed summary of activities involved in these phases are explained as thus;

Phase I: Modification of SET functions for unstructured data (ESET1)

Sequential Exception Technique (SET) involves finding anomalies in categorical data from log files of large databases. In contrast to the existing or traditional SET, this study proposes to modify SET to accommodate textual data since text is a reliable source of meaningful information and knowledge. However, series of tasks were performed to modify SET functions into ESET, these tasks include; extraction of text from data source (UCI machine learning repository), text-pre-processing tasks like lamentization, stop-word removal and stemming were also performed. Afterwards, the performance of cardinality and identification of useful text from documents. It was noticed that the traditional SET makes use of variance to compute deviants which may not be suitable enough for text data as explained in (Chapter 1, sub-section 1.3). In the case of the ESET, a better similarity measure was adopted namely the optimized cosine similarity. Moreover, instead of leveraging the cardinality and smoothing functions used in the traditional SET to count and smoothing detected anomalous data into desired target data, count-vectorization was adopted in ESET to count vectors and prune text document into desired form by employing the minimum and maximum difference functions.

Phase II: Hybridizing ESET1 with Latent Semantic Analysis for semantic-based anomaly detection (ESET2).

This phase allows a more profound means of relating human judgments for semantic similarity between words. This was aimed at understanding the consequences of the overall word-based similarity in documents. It is imperative to note that, the similarity/ dissimilarity estimates determined by Latent Semantic Analysis (LSA) are not simple contiguity frequencies, co-occurrence counts, or correlations in usage. However, LSA exclusively rely on a ground-breaking mathematical analysis known as Singular Value

Decomposition (SVD) (Chang, Yih, and Meek, 2013; Kiriti1, 2017; Steinberger and Ježek, 2004). SVD ventures the dimensionality of text data by pinpointing the space spanned by the left singular vectors. $X = U\ x\ S\ x\ V^T$ where U is a $(u\ x\ n)$ matrix, S is a $(n\ x\ n)$ matrix and V^T is a $(v\ x\ n)$ matrix with n 'latent semantic' measurements. The matrices U and V relates to terms and documents in the new space contains the left and right singular values of X, individually, and the diagonal of S contains the singular values of X in decreasing order. A decrease of the original matrix into k-measurements can be performed by taking the k largest singular values. This phase aims at analysing identified semantics.

Phase III: Integration of Word Sense Disambiguation algorithms with ESET2 (Lesk and Selectional preference) (ESET3). It was also noticed that in the previous phase, more operation needs to be performed to resolve issues relating to text ambiguity since ESET2 with Latent Semantic Analysis may not be a good fit for resolving polysemy and synonymous issues in words. Canonical word order is a multilevel grammar that captures the semantics in text. It does this by equating the subject in an underlying sentence structure with the semantic role of agent and its stipulated role remains unchanged as transformation is in process to derive word order. The benefits are to have a solid source of knowledge generated by concept definitions and dictionaries. Another important question that is brought up is why canonizing vector length and what are its importance in identifying text semantics? The reason was to resolve ambiguity issues relating to synonym and polysemous terms found in documents.

As earlier stated, LSA was unable to tackle properly text ambiguity problems. This is because, senses found in text documents cannot be properly disambiguated from LSA

Universiti Utara Malaysia

operations only. More operations need to be considered when disambiguating text document. Therefore, this study introduces text canonization by combining two WSD algorithms namely Lesk and Selectional preference for word sense disambiguation. These combined algorithms with the Enhanced Sequential Exception Technique. This enables easy and effective detection of semantic-based text anomaly in documents.

Phase IV: Visualization semantic-based text anomaly using FOL and concept network graph

Detecting semantic based anomalous text in document is not complete if it cannot be physically represented or interpreted. The study made use of First Order Logic (FOL) to draw a formal expression of text information completeness in documents. Terms are matched with the context of idea conveyed in corpuses. If a context is seen to have a meaningful and complete idea, then it is apparent that knowledge can be shared on that context. Then again, if knowledge is not shared or no meaning is gotten from context, text information is said to have an incomplete meaning. FOL rule recognizes terms like individual variables, individual constants and predicates which takes differing forms, these forms are then connected using edges and nodes based on relatedness with the use of a Concept Network Graph to represent illustration of the detected semantic based text anomalous data or identified knowledge from document.

3.3 Research Data

Four data with varying domain fields were mainly experimented upon in this study. These includes News, Journal articles and Business organization related data. However, the aim

of exploring these data is to give a wider picture of how efficient the ESET can be applied on other data in varying domain fields for future research purposes.

ENRON data was original sourced by David Newman from the University of California, Irvine. It weighs almost 385MG as compressed file and 1.32GB as unzipped file (Gloor et al. 2006; Kumar Palaniswamy Supervisor & Aldous, 2015; Zhou et al. 2010). ENRON data was noticed to have a combined form of dictionary, where each key-value pair in the dictionary corresponds to names of staff and value was termed as another dictionary. Moreover, features included in ENRON data can be divided into three categories, salary features, stock features and email features. It was also observed that the identified data type in ENRON can be categorized into Numerical Data (Salary, number of emails) Categorical Data (Job Title), Time Series Data (Timestamps on email), Text Data (contents from emails, To / From fields email) (Koehrsen, 2017). This research focuses mainly on text data (contents from emails). The possible anomalous text that could be detected in these data includes; Outrageous or abnormal total payment information, Persons of Interests (POI), closeness or degree of Mail messages to and from the POI to identify anomalous or suspicious departments.

Neural Information Processing System (NIPS) is a non-profit making organization whose aim is to foster academic research on Neural Information Processing Systems. NIPS benefits from combined domain of mathematical, biological and computational sciences. Conference on Computational Neuroscience randomized 1500 documents with 6.4 million total words and 12,419 unique words. NIPS data contains three important documents namely the paper.csv, authors.csv and authors_paper.csv (Goodfellow, 2016). Possible anomalous text that can be identified in this data includes; Authors with

Universiti Utara Malaysia

outstanding numbers of publications, frequency of text content and title from published journals or the diverting difference in text information content amongst journal abstracts.

Twenty News Groups data is approximated to have 20,000 newsgroups documents (Ren & Sohrab, 2013). This news data are partitioned evenly to 20 varying newsgroups, which was originally collected by Ken Lang, for his 'Newsweeder: Learning to filter net newspaper'. Daily Kos Blogs Data is an American political blog that publishes political news and opinions, typically adopting a liberal stance. Random set of blogs taken from their website. 3430 documents,6906 unique words and approximately 467714 total words in it. Frequency of text from blog could be identified as anomalous data. Some sample data like (MS paraphrase test corpus) were used in this study just to evaluate performance of the technique with existing studies.

3.4 Experimental Design

This section serves as a guideline within which the study experiment was conducted. Respective functions performed in each phase were described for easy understanding of the technique as thus; In Phase 1, Modified SET was integrated and tested with other similarity functions like Manhattan, Euclidean and Cosine similarity to know which function produces better results in identifying text anomalies through the detection of similar and dissimilar text. It was noticed that the optimized cosine function with the modified SET produces a better result and was used to formulate ESET1. ESET1 was then benchmarked with existing works by (Gloor et al. 2006; Hardin et al. 2015) to evaluate performance of ESET1 with existing studies.

Phase 2 made use of LSA model to analyse the semantics of identified text anomalies.

Models like LDA, HDP, LSA and NMF were tested with ESET1. However, ESET1 integrated with LSA (ESET2) proved to have a better coherence and consistency measure in topic or concept identification from text documents. This will be shown and described in the next chapter. ESET+LSA was evaluated with an existing study performed by Yin & Wang (2016) whose research was aimed at detecting semantic based text anomalies.

In Phase 3, words were canonized for sense disambiguation purpose in text documents. To perform this, both knowledge and corpus based WSD approach needs to be properly understood. These approaches were reviewed and compared in this phase and it was noticed that combining both approaches would yield better sense disambiguation accuracy. A combined WSD algorithm using Lesk and Selectional preference was performed to disambiguate senses from text documents. The combined WSD algorithms was noticed to perform almost like human judgement on identifying similar and dissimilar sentences. To further this study, ESET2+ combined WSD algorithms (ESET3) was compared with the work of (Li, et al. 2009) to test for sentence semantic similarity and was also tested with (Mahapatra et al., 2012) to detect diverging topics in documents. Overall, ESET3 was able to resolve issues relating to ambiguity (synonyms and polysemous) in words and as well gave meanings to the detected anomalous terms by identifying diverging side information in text documents.

Phase 4 represents the identified semantic-based text anomalies with the FOL and Concept Network Graph for better understanding and interpretation of predicates and arguments in text documents. An illustration on how the research experiments were performed is shown in Figure 3.2

Universiti Utara Malaysia

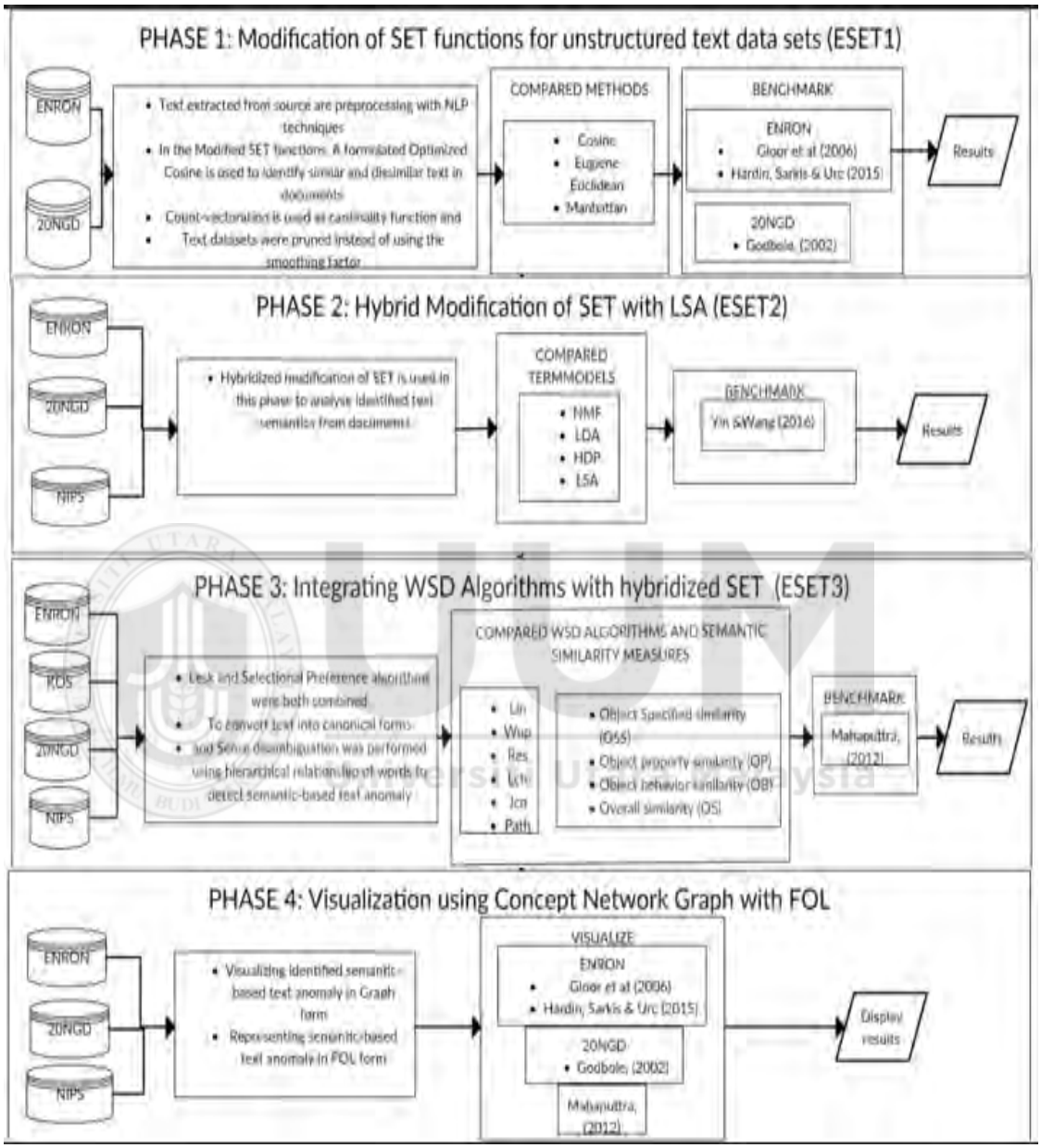

Figure 3.2. Research design of ESET for semantic-based Anomaly Detection

This chapter provides a detailed description of the research methodology from the identified phases to know the kinds of activities that were performed, and the outcomes of each phases were also shown in Table 3.1. for clearer understanding of the study.

Table 3.1.

Experimental design phases with expected outcome

PHASE	STEPS	ACTIVITIES	OUTPUT
Phase 1 RQ/RO 1 and 2 (ESET1)	SET SET + text pre	• Find anomalous text data with (SET functions). • Prepare SET for textual data	• Reduce text sparsity • Calculate similarity and Dissimilarity of text to find anomalous data
	Optimize Cosine similarity ESET1	Formulate Optimized Dissimilarity and Similarity measures	• Modify cosine similarity function with ESET to similar and dissimilar text as anomalous text
Phase 2 RQ/RO 3 (ESET2)	ESET1+LSA	ESET1 with Latent semantic Analysis model using Single Value Decomposition	• Analyse and Detect semantic-based text anomaly
Phase 3 RQ/RO 4 (ESET3)	ESET2+WSD	• ESET2 was tested with other semantic similarity measures to evaluate and compare their performance • ESET2 with combined WSD algorithms namely Selectional preference and Lesk algorithm.	• Text normalization into its canonized form • Resolve text issues relating to synonyms and polysemy for better sense disambiguation • Detect semantic-based text anomaly
Phase 4 RQ/RO 5	Concept Network Graph with FOL	Concept Network Graphs First order Predicate Logic	• Resolve predicate and argument issues found in text • Reliably represent text semantics

Universiti Utara Malaysia

3.5 Evaluation Measures

Common machine learning techniques includes supervised, unsupervised and anomaly detection approach. However, each approach calls for different performance evaluation methods. The lack of labels on an unsupervised learning models for training data makes evaluation problematic. This is because there are no models to compare meaningfully the explained outcomes or results of data. In this case, manual grouping of data with algorithms is not an option due to labour and time constraints (Rosenberg & Hirschberg, 2007). A more reliable way to determine how well algorithms performs is needed. An unsupervised approach is seldomly used in the context of a more complex workflow, in which an extrinsic performance function can be defined. However, it is plausible to create an external data by hand-labelling to test for accuracy (so-called gold standard). The evaluation of ESET for anomaly detection was performed empirically. The purpose of empirical evaluation was to measure the accuracy of semantic-based anomalous text data detected (Powers, 2015; Provost et al. 1997).

Many metrics such as F-measures, Recall, and Precision have been explored to detect classification error rate in text documents. To measure the accuracy of information extractor, basic measures for text retrieval includes the precision and recall. Precision is said to be the proportion of retrieved material that is relevant, while recall is the proportion of relevant material retrieved (Powers, 2015; Provost et al. 1997). These can only be measured if text anomalies are known. Using the above parameters, the measurement of the experiments was formulated. F-measure combines precision and recall scores as shown in Table 3.2.

Table 3.2.

Confusion Metrics for a two-class classifier

		Actual value	
		Positive (1)	**Negative (0)**
Predicted Value	Positive (1)	*A*	*B*
	Negative (0)	*C*	*D*

Source: (Godbole, 2002)

Definition of parameters used in confusion metrics Table 3.1 alphabet *ABCD* are termed with specified meaning. *"(A = TP)* it is the number of correct predictions that an instance is negative*, (B = FP)* it can be defined as the number of incorrect predictions that an instance is positive, *(C = FN)* it can be defined as the number of incorrect predictions that an instance is negative and *(D = TN)* is the number of correct predictions that an instance is positive" (Powers, 2015; Provost et al., 1997).

The following is the measurement for precision and recall;

- *Precision = (|relevant ∩retrieved|)/(|retrieved|) or* Precision *[TP / (TP+ FP)]* (3.1)
- *Recall = (|relevant ∩ retrieved |) / (|relevant|) or* Recall = Sensitivity = Total Positive Rate is a proportion of cases that were correctly identified as positive. It is defined as *[TP / (TP + FN)] = [D / (C + D)]* (3.2)
- Accuracy is defined as the portion or part of the sum number of predications that is correct. It is given as *[(A + D) / (A + B + C + D)] or [(TP + TN) / (TP + FP + FN + TN)]* (3.3)
- False Positive Rate are the proportions of cases that were incorrectly identified or classified as positive. *[B / (A + B)] or [FP/ (TP +FP)]* True Negative *[TN / (TN + FP)]* False Negative *[C / (C + D)]* (3.4)

$$F\text{-measure} = 2 \times [(precision \times recall) / (precision + recall)] \quad (3.5)$$

Universiti Utara Malaysia

In an unsupervised machine learning approach, it is important to satisfy some performance evaluation metrices like; completeness, homogeneity and v-measures. To satisfy homogeneity criteria, a clustered data must assign only those data points that are members of a single class to a single cluster. That is, the class distribution within each cluster should be skewed to a single class, that is, zero entropy. To determine how close a given clustering is done by examining the conditional entropy of the class distribution given the proposed clustering. In the perfectly homogeneous case, this value, *H(C|K),* is 0. However, in an imperfect situation, the size of this value, in bits, is dependent on the size of the data and the distribution of class sizes. *H(C|K)* is 0 when each cluster contains only members of a single class, a perfectly homogenous clustering (Rosenberg & Hirschberg, 2007).

Completeness is symmetrical to homogeneity. To satisfy the completeness criteria, a clustering must assign all those data points that are members of a single class to a single cluster. Lastly, V-measure is an entropy-based measure which explicitly measures how successfully the criteria of homogeneity and completeness have been satisfied. V-measure is computed as the harmonic mean of distinct homogeneity and completeness scores, just as precision and recall are commonly combined into F-measure (Rosenberg & Hirschberg, 2007). Mathematical Definition of Unsupervised Machine learning performance evaluation. Thus, to adhere to the convention of 1 being desirable and 0 undesirable, homogeneity is defined as: $h = \left\{1 - \frac{H[CK]}{H(C)}\right\}\ if\ H(C,K) = 0$ where $H(\vdash c\dashv|k) = -\sum\_(\sum k = 1)\hat{}|k|\sum\_(c = 1)\hat{}|c|\ \llbracket ack/N\ log\ \llbracket ack/(\sum\_(c = 1)k\_ac\hat{}1\ k)\rrbracket\rrbracket$ (3.6)

Universiti Utara Malaysia

$$H(c) = -\sum\_(c = 1)\^|c|\ \llbracket(\sum\_(k = 1)\^|k|\ \llbracket a\_c\ k\rrbracket\ )/n\ log\ \ \llbracket(\sum\_(k = 1)\^|k|\ \llbracket a\_c\ k\rrbracket\ )/n\rrbracket\rrbracket \quad (3.7)$$

Completeness is defined as: $h = \left\{1 - \frac{H[CK]}{H(C)}\right\}\ if\ H(C,K) = 0$ where (3.8)

$$H(\vdash k\dashv|c) = -\sum\_(c = 1)\^|c|\sum\_(k = 1)\^|k|\ \llbracket ack/N\ log\ \ \llbracket ack/(\sum\_(k = 1)k\_ac\^1\ k)\rrbracket\ \rrbracket \qquad H(k) = -\sum\_(k = 1)\^|k|\ \llbracket(\sum\_(k = 1)\^|c|\ \llbracket a\_c\ k\rrbracket\ )/n\ log\ \llbracket(\sum\_(k = 1)\^|k|\ \llbracket a\_c\ k\rrbracket\ )/n\rrbracket\rrbracket \quad (3.9)$$

Based upon these calculations of homogeneity and completeness, then a V-measure solution was weighted by computing harmonic mean of homogeneity and completeness

$$V_\beta = \frac{(1+\beta)*h*C}{(\beta*h)+C} \quad (3.10)$$

Similar to the familiar F-score or F-measures, if $\beta > 1$, then completeness is weighted more strongly in the calculations, If β is < 1, then Homogeneity is weighted more strongly. Measurement scores on tested models cannot be seen in the Silhouette Coefficient performance evaluation, it is a quite small measurement. Other measures like Adjusted Rand Index and V-score are based on theoretical information of evaluated scores (Rosenberg & Hirschberg, 2007). Another interesting unsupervised machine learning approach is modelling terms. Terms, Topics or Concepts are sets of important words automatically from unlabelled documents in an unsupervised manner.

This is an attractive method that brings structure to otherwise unstructured text data, but topics or concept are not guaranteed to be well interpretable, therefore, coherence evaluation measures have been employed to distinguish between relevant and irrelevant terms. Text coherence can be evaluated both intrinsically and extrinsically. Intrinsic Measure is represented as UMass which measures and compares word only to the preceding and succeeding words respectively, ordered word sets are needed for this kind

Universiti Utara Malaysia

of measures. More so, it uses pairwise score function as the empirical conditional log-probability with smoothing count to avoid computing the logarithm of zero. On the contrary, the Extrinsic Measure is represented as UCI. In UCI measure, every single word is paired with every other single word. The UCI coherence uses Pointwise Mutual Information (PMI). Both Intrinsic and Extrinsic measure compute the coherence score *c* (sum of pairwise scores on the words *w1, ..., wn* used to describe the topic).

Research experiment in this study was performed empirically, which was solely aimed at achieving a significant level of accuracy.

3.6 Summary

In addition, unsupervised approach was mostly used in the context of a more complex workflow, in which an extrinsic performance function can be defined. The goal for the empirical evaluation was focused on measuring the accuracy of ESET using varying evaluation measures. These measures are determined mainly by the benchmarked studies.

The scientific experimental approach was the basis for the methodology as new methods for semantic-based text anomaly detection was developed. The chapter began with a description of steps in the experimental design and discussion on evaluation measures was presented with its experimental settings. To further the explanation of the discussed ESET research design, a detailed-out experimentation of ESET with data were analysed in subsequent chapters (4,5,6 &7). This is aimed at achieving the study objectives and to provide some contributions for future studies with similar research objectives.

Universiti Utara Malaysia

CHAPTER FOUR

MODIFICATION OF SET FUNCTIONS FOR UNSTRUCTURED TEXT DOCUMENT (ESET1)

4.1 Overview

This chapter gives an elaborate analysis of ESET1 on text data as shown and described in (section 3.2) of the study. Subsequent sections in this chapter illustrates ESET steps, methods and result findings from experimental performance with benchmarked study.

4.2 Introduction

In this chapter both frequent and infrequent text data were considered as anomalies in documents. Text similarity measure is the core of many techniques in text mining and it plays a pertinent role by delivering how likely terms share same or deviating ideas. This is because different algorithms merely make decision on what identity is associated with every term to form useful ideas in documents. The goal of text mining is to mine useful text either meaning or surface closeness (semantics and lexical similarity). In the context of this study, which aims at detecting semantic-based text anomalous data in documents, both lexical and semantic text similarity are to be considered for the study objectives. Thereby, considering the granularity from character, word, phrase, sentence and document level as well making preference to understand the most suited metrics for detecting semantic based text anomalies.

ESET1 was divided into stages in this chapter, these stages are poised on performing

Universiti Utara Malaysia

functions which are tailored towards achieving the research objectives of the study as will be further explained in subsequent sections. The stages in this study was aimed at enhancing SET to ESET1 by following a multi-facet step as thus;

- SET was used to identify anomalies in form of POI from ENRON data with highest total payment information.
- ESET1 (Text-pre-processing + Modified SET) was used to extract information relating to identified POI mail message.
- ESET1 (Text-pre-processing + Modified SET) ESET1 on 20NEWSGROUPS data to detect the most anomalous groups based on lexical word relationship

The aim of breaking ESET1 into stages was to make sure that the ESET1 can detect text anomalous data as used in this study.

4.3 Performing Sequential Exception Technique (SET) on ENRON data

In this section, the traditional SET was experimented upon to detect text anomalies in documents. SET was originally used for categorical data, both on the contrary in this study, SET was employed to detect POIs with outrageous total information on bonus, payment of salary and stock data. Some existing projects performed similar task, using feature selection algorithms to identify POIs on ENRON data (Koehrsen, 2017; Leyzerov, 2017; Pappas, 2018). Consequently, SET was able to identify similar outlying or anomalous data as compared with the existing projects. This was simply done by computing the variance score of the total payment information of each ENRON staffs and then using smoothing factor to identify the most outrageous payment information as shown in Table 4.1.

Table 4. 1.

Persons of Interest outlined queries

Queries	SET Snippet output of payment information for identified POIs
What is the highest value of stock options found for some POIs	'SKILLING JEFFREY K'=26093672 'KEAN STEVEN J'=6153642 'LAY KENNETH L=49110078 'WHALLEY LAWRENCE G'=6079137 **'LAY KENNETH L=49110078**
ENRON POI with the highest total payment?	'SKILLING JEFFREY K'= 5600000 'KEAN STEVEN J'= 1000000 'LAY KENNETH L'= 7000000 'WHALLEY LAWRENCE G'= 3000000 **'LAY KENNETH L'7000000**
Anomalous POI total messages received?	'FROM_THIS_PERSON_TO_POI'=16 'FROM_MESSAGES' =36 **'FROM_POI_TO_THIS_PERSON =123**

From Table 4.1 it was deduced that data points like Kenneth Lay which possess the highest total payment information on bonus, salary and stock may have other interesting information in text form. Since Kenneth Lay was once an active CEO in ENRON company, extracting information from him may lead to interesting information regarding other persons and departmental activities in the company. Judging from this fact, mail messages related to Kenneth Lay needs to be extracted and analysed using ESET1.

However, this study is not interested in detecting scandalous or fraudulent activities in ENRON company. The aim of the study is poised towards showing that ESET can be applied or employed to identify knowledge for improved decision making and information sharing purposes from text documents by detecting anomalous text information.

Universiti Utara Malaysia

4.4 Enhanced Sequential Exception Technique (ESET) for Text data

In this study, traditional SET was improved as modified SET to fit in textual data instead of categorical data. More so, the study aims at identifying information and how detect similar and dissimilar text data as shown in below

Step1: Import libraries and extract text mail messages and pre-process them using Natural Language Processing (NLP) techniques.
Step 2: The stop word is compared to target text in form of array using sequential search technique. If it matches, the word in array is removed, and the comparison is continued till length of array. After removal of stop-word completely, another stop-word is read from stop-word list and again algorithm follows step 2. The algorithm runs continuously until all the stop-words are compared.
Step 3: Resultant text devoid of stop-words were displayed and fed as a list.
Step 4: Create document term- matrix (using count vectorization or tf_idf) to reduce sparsity also to convert sparse matrix into data frames so cardinality or word frequency can be easily computed.
Step 5: Use count-vectorizer or *tf-idf* to prune text into preferred or desired numbers
Step 6: Words in data frames are then compared to check similar and dissimilar text and
Step 7: Dissimilarity and cardinality function to identify numbers of similar and dissimilar text in the list.
Step 8: Calculate results by comparing relevance with retrieved results to compute precision recall and F1-score

*Figure 4. 1.*Steps in detecting dissimilar /similar text using ESET

The necessary steps in performing ESET1 are shown. However, these processes posed many challenges, especially during extracting and pre-processing text data. One of the most challenging tasks faced was text sparsity issues in document. A document term matrix was created to reduce issues relating to text sparsity. However, it makes sense to use only non-zero values to perform operations as zero multiply by any value will result to zero and may not give better accuracy in results. This can be resolved by converting text into data frame, which allows the performance of a better computation of similarity and dissimilarity identification in text data. However, it is most suitable for results to be

displayed using term frequency, this helps to down weight words that occurs frequently across documents.

4.4.1 Optimized cosine

Cosine similarity seems to be the most meaningful text similarity measure, because it considers the relative frequency of words instead of the actual frequency. Take the case where there are two articles, *A* and *B*, and article *A* is the same as article *B,* except each word in *A* appears twice as many times in *B*. The similarity measure ought to indicate the articles are highly similar. The Eugene similarity would be 0.5, cosine similarity would be 1, and Manhattan similarity would be some non-zero number. With Jaccard and *L2* similarity, the number of words in each article has some influence on the similarity measure, so when one article has a lot more words than another, they will appear more dissimilar.

Universiti Utara Malaysia

Cosine similarity is a building block to other similarity measures like Pearson correlation coefficient, Ochiai coefficient and Levenshtein distance. It assumes that text documents are vectors which is enough to capture reasonable information from original text documents. Vectors are counted in this study using the count-vectorizer. This is like the operation of a cardinality function in traditional SET. Count-vectorizer computes maximum difference and minimum difference to prune data into desired size. This can be also seen as smoothing text into desired size. The maximum difference is also known as corpus specific which is used to remove terms that appears too frequently. On the contrary, minimum difference removes terms that appears too infrequently and vice-versa. Based on the identified problems, this study proposes the formulation of an optimized cosine

with ESET following the steps shown in Figure 4.3.

Step1. Identify K-docs in the collection nearest to the term ⇒K largest query d Step 2. Equalize dimensions of vectors Step 3. Speed up vector space ranking Step 4. Formulate dimension value or pseudocodes for optimized cosine similarity for ESET1 dissimilarity function, Perform data *I* as Input Output Max value of Input, where number is same as the */I/* cardinality For every doc length is initialized, For every term query of terms Divide scores doc by doc length, Return Top-*K* components of scores [] Step 5 ***Similarity*** **(*t, t*^{1}) = *doct* = *[∑ sim (t• t*^{1}*) doct*$^{1}_{i}$*] / [∑ sim (t, t*^{1}*)]** ***D*issimilarity (*t, t*^{1}) = *doct* = 1- *[∑ sim (t• t*^{1}*) doct*$^{1}_{i}$*] / [∑ sim (t, t*^{1}*)]**

Figure 4. 2. Optimization of cosine function

Optimized cosine in this case cannot be over emphasized, nevertheless, *D*issimilarity (*t, t*^{1}) = *doct* = 1- *[∑ sim (t• t*^{1}*) doct*$^{1}_{i}$*] / [∑ sim (t, t*^{1}*)]* is to consider cosine function as dissimilarity measure. This is mostly applied on the bag of words data. While the

Similarity (*t, t*^{1}) = *doct* = *[∑ sim (t• t*^{1}*) doct*$^{1}_{i}$*] / [∑ sim (t, t*^{1}*).* (4.1)

Similarity is mostly applied on full text document

(a) Performing ESET1 on ENRON data

ESET1 was initially tested on ENRON mail messages to see how the ESET would perform on large text document. However, before it can be applied on ENRON mail, some text-pre-processing task needs to be performed as illustrated in figure 4.3.

Step1. Analyse email messages as input files Step2. Extract email body of most To and From Kenneth lay emails messages then append each mail. Step 3. The extracted mail body are then parsed as well appended Step 4. Appended files are then written into the root directory file of systems Step 5. The numbers of list of email sent to and received from messages with email addresses are then displayed on the written files saved in the root directory of the system.

Figure 4. 3. Parsing extracted mail messages

Parsing and counting approach are applied on Kenneth Lays topmost sender and receivers as shown in Figure 4.4. This was further detailed by showing the list of POIs names.

Top email senders and receivers in all Emails:
To:[('richard.shapiro@enron.com',15149),('jeff.dasovich@enron.com',14207),
('tana.jones@enron.com',12828),('steven.kean@enron.com',12754),('sara.shackleton
@enron.com',11433)]
From:[('kay.mann@enron.com',16735),('vince.kaminski@enron.com',14368),('jeff.d
asovich@enron.com',11411),('pete.davis@enron.com',9149),
==========
Top email senders and receivers in K lays Emails:
To:[('kenneth.lay@enron.com',2039),('klay@enron.com',1903),
From:[('rosalee.fleming@enron.com',856),('brown_mary_jo@lilly.com',82),('leonar
do.pacheco@enron.com',78'tori.wells@enron.com',58)

Figure 4. 4. Extracted top POIs mail messages from senders and receivers

To know the performance of the ESET, ESET1 was performed on thirteen top mail messages senders and receivers identified from Kenneth Lays mail messages. This was aimed at knowing POI that share similar text information with Kenneth Lay. To know the most information shared, three similarity measures were compared with ESET1 as shown in Table 4.2.

Universiti Utara Malaysia

Table 4. 2.

Comparing similarity/ dissimilarity measure of ENRON identified POIs

Names of POI	ESET Cosine	ESET Eugene	ESET Manhattan	ESET Cosine	ESET Eugene	ESET Manhattan
Similarity score				**Dissimilarity score**		
Ken-lay	1.00	0.00	1.00	0.00	1.00	0.00
J-Skillings	0.86	0.51	0.73	0.24	0.49	0.27
R-Shapiro	0.87	0.52	0.71	0.23	0.48	0.29
K-Mann	0.70	0.77	0.70	0.30	0.23	0.30
J-Dasovich	0.79	0.65	0.71	0.21	0.35	0.29
T-Jones	0.57	0.93	0.70	0.43	0.07	0.30
S-Kean	0.84	0.56	0.73	0.26	0.44	0.27
S-Sara	0.68	0.80	0.70	0.32	0.20	0.20
J-Steffes	0.76	0.70	0.64	0.24	0.30	0.30
M.-Taylor	0.70	0.79	0.71	0.30	0.21	0.29
D-Pete	0.53	0.87	0.68	0.47	0.13	0.31
Chris -G	0.62	1.00	0.66	0.38	0.00	0.34
K-Symes	0.58	0.92	0.67	0.42	0.08	0.33

Universiti Utara Malaysia

To further explain this study, scores from Table 4.2 were compared with an existing study by Gloor et al.,(2006) to show the potential suspicious POIs. It was apparent that ESET+cosine similarity exhibited better results compared to other similarity measures. It is also important to note that limited list of POIs was identified due to the size of information in proportion with the running power of system used for this study. Figure 4.5 shows POIs with closest information with Kenneth Lays mail messages using the result generated from ESET+Cosine (ESET1).

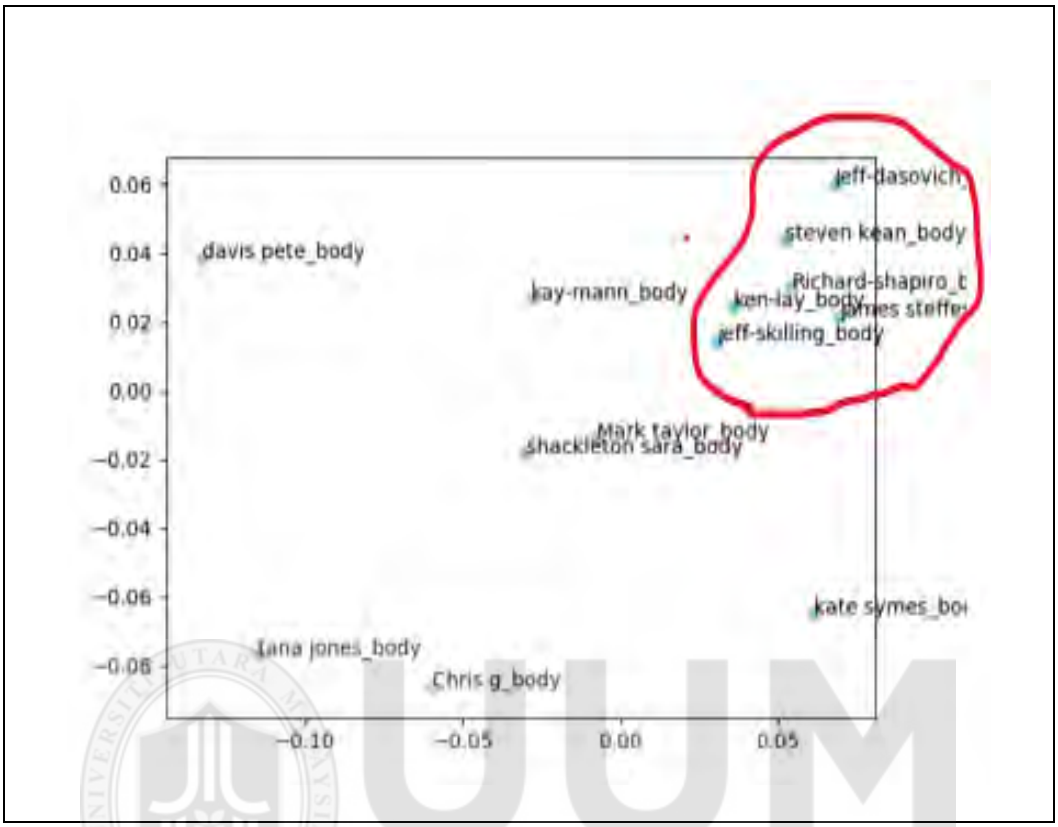

Figure 4. 5. POIs message similarity

Universiti Utara Malaysia

To compare information centrality, relatedness of identified POIs with Kenneth Lay, an online whitepaper list of ENRON staffs with designated departments created by the University of Edinburgh with another study proposed by (Hardin et al., 2015) was used as a backup information for comparison. The relatedness of POIs mail messages with Kenneth Lays as shown in Figure 4.5. was compared with the study of Gloor et al.,(2006) who identified or labelled some staffs as potential suspects in ENRON company using temporal link analysis. To evaluate the performance of the ESET, precision and recall using relevance and retrieved text results from ESET were computed

Table 4.3 showed that ESET was able to match results with six of the identified POI in

comparison to the compared study. Only one POI was unable to be identified (Robert Badeer). Robert may have been communicating offline which possibly cannot be noticed by the ESET. In other words, it can be deduced that 6 relevant POIs were identified from the retrieved 7 POIs. The aim is to get the F-1 score which also captures accuracy and sensitivity of anomalous data detected.

Table 4. 3.

Comparing POI names with identified departments

Suspicious departments in ENRON from Gloor et al.,(2006)	**(Relevance)**	**(Compared with top mail message retrieved from Kenneth lays results)**	**(Retrieved)**
Former CEO Jeff Skillings	Former CEO	Jeff skillings	1
CEO, Enron Energy Kenneth Lay	CEO,	Kenneth Lay Yes	1
V.P Government affairs James steffes	V.P Government	James steffes	1
Employee Director	Employee Jeff Dasovich, Mark Taylor, Chris Germany, Kay Mann	Jeff Dasovich, Mark Taylor, Chris Germany, Kay Mann Not related	1
Vice President Regulatory Affairs Richard Shapiro	Regulatory Affairs Richard Shapiro	Richard Shapiro	1
VP and Chief of Staff	VP and Chief of Staff Steven Kean	Steven Kean	1
Manager	West	Robert Badeer	0

Universiti Utara Malaysia

Power/ Mgr Trading	

Table 4.4 showed some active departments identified by detected POIs except for the West Power Trade Managing department. A precision score of 86% was recorded, 100% recall score and 90% score of accuracy. Further exploratory analysis can be made on this data to improve results and as well effectively identify useful information from anomalous data in documents.

Table 4. 4.

Results of ESET1

Metrics	**Scores (%)**
Precision *[(Rel +Ret)/Rel]*	0.86
Recall *[(Rel +Ret)/Ret]*	1.00
F1-Score *2 X [(Prec X Rec)/(Prec + Rec)]*	0.90

Universiti Utara Malaysia

(b) Performing ESET1 on 20Newsgroups (20NG) Data

To further explore ESET1, topics or group of documents were also identified from 20NG. 20NG has a post of 20 topics, when comparing each article with every other articles, it was noticed that some articles are closely related like comp.sys.ibm.pc.hardware / comp.sys.mac.hardware, while others are highly unrelated like misc.forsale / soc.religion.christian. Table 4.6 shows list of the 20 newsgroups, partitioned (more or less) according to subject matter.

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

Figure 4. 6. 20NG topic grouping

ESET is used to identify topics that has most similar text with other topics and the topic with the most dissimilar text. Figure 4.7 shows some newsgroups have similar demographics. Other similar pairs include soc.religion.christian+alt.atheism and soc.religion.christian + talk.religion.misc. This seems plausible, that there is some overlap discussing religion or lack of it.

Universiti Utara Malaysia

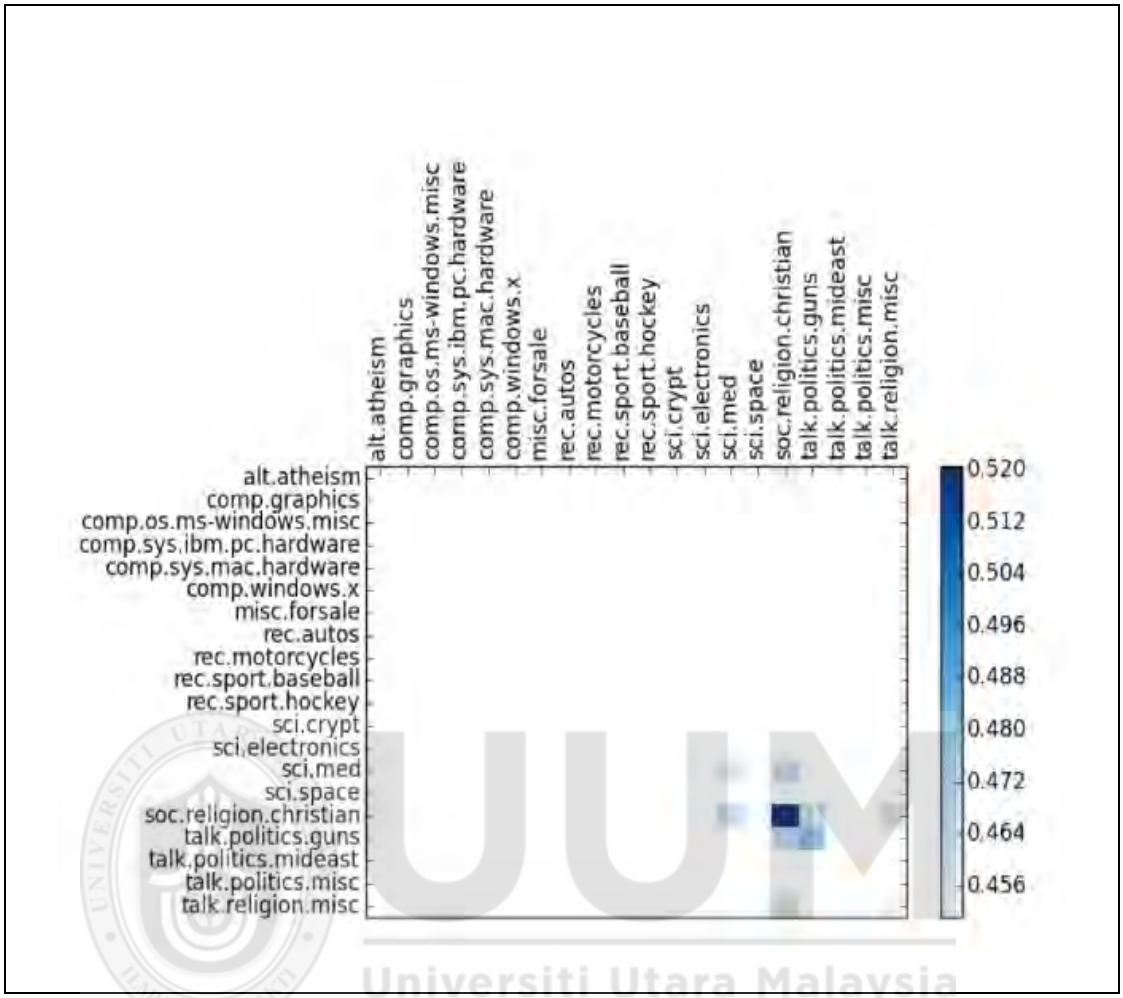

Figure 4. 7. ESET+Cosine 20Newsgroups with similar themes (religion)

Universiti Utara Malaysia

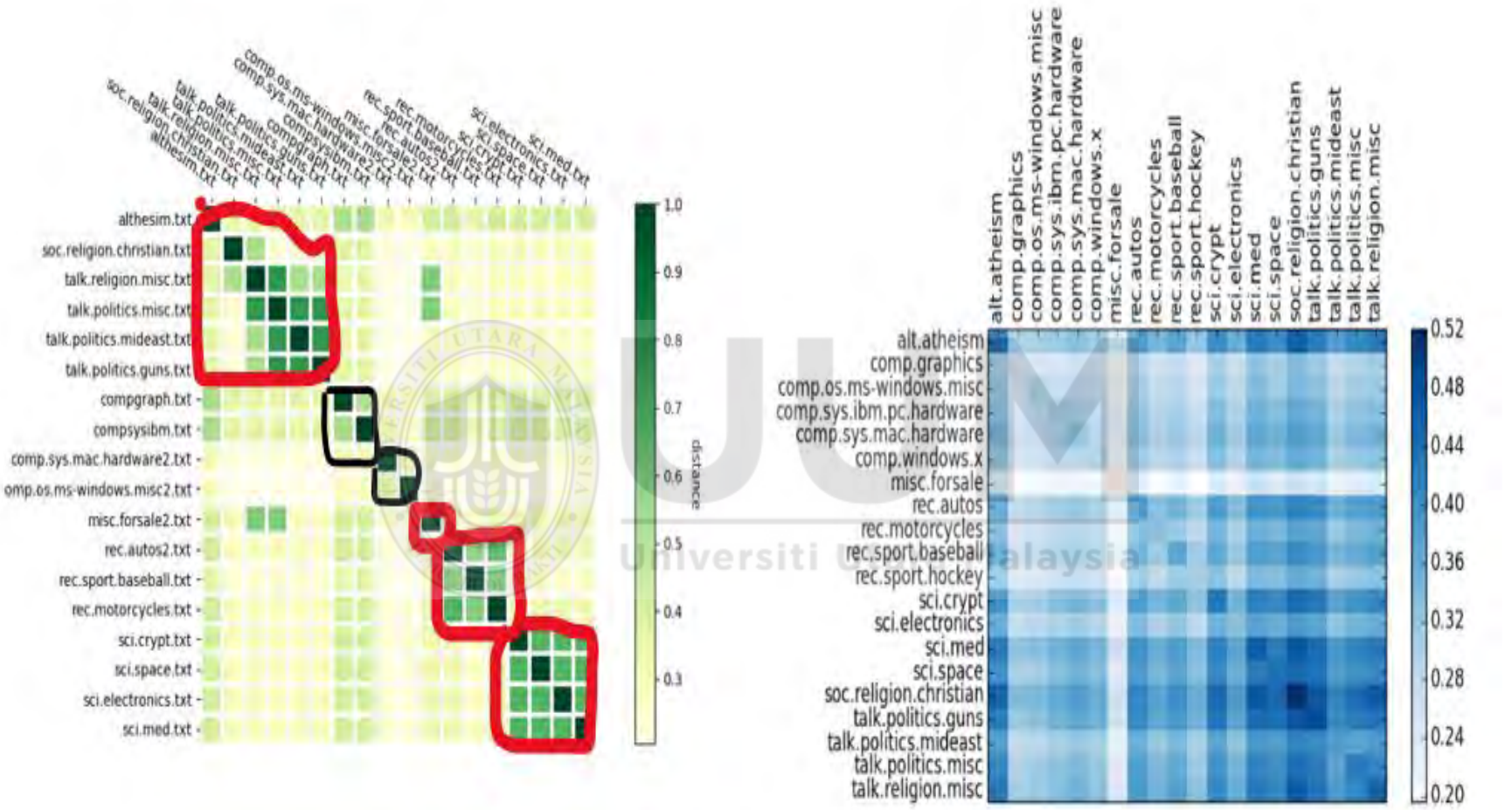

Figure 4. 8. ESET+Cosine

Figure 4. 8. showed that the topic Misc for sale is entirely different from other topics. This shows that textual contents from topics from different groups are unrelated to other topics.

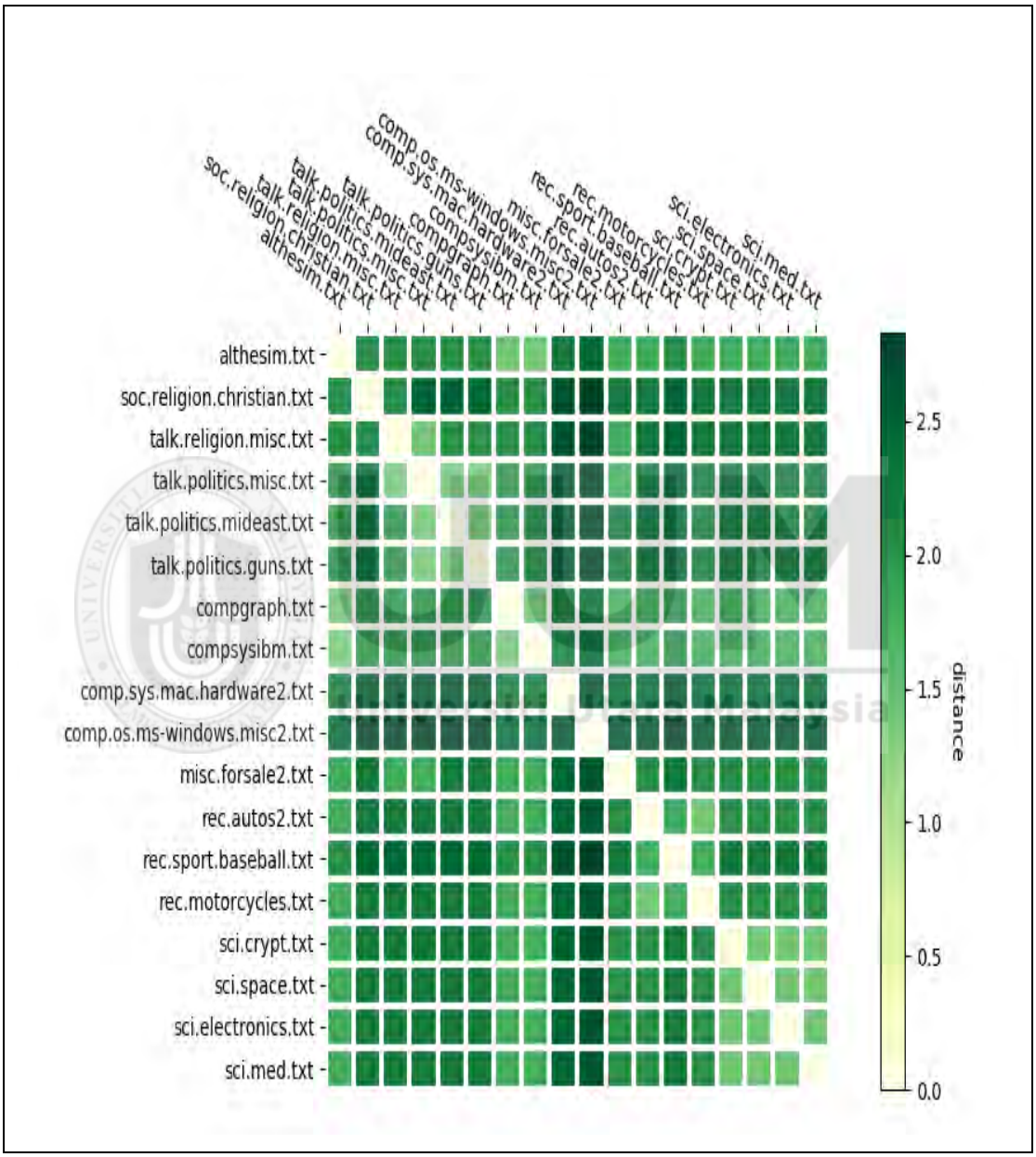

Figure 4. 9. ESET+Eugene with marks indicating similar and dissimilar groups

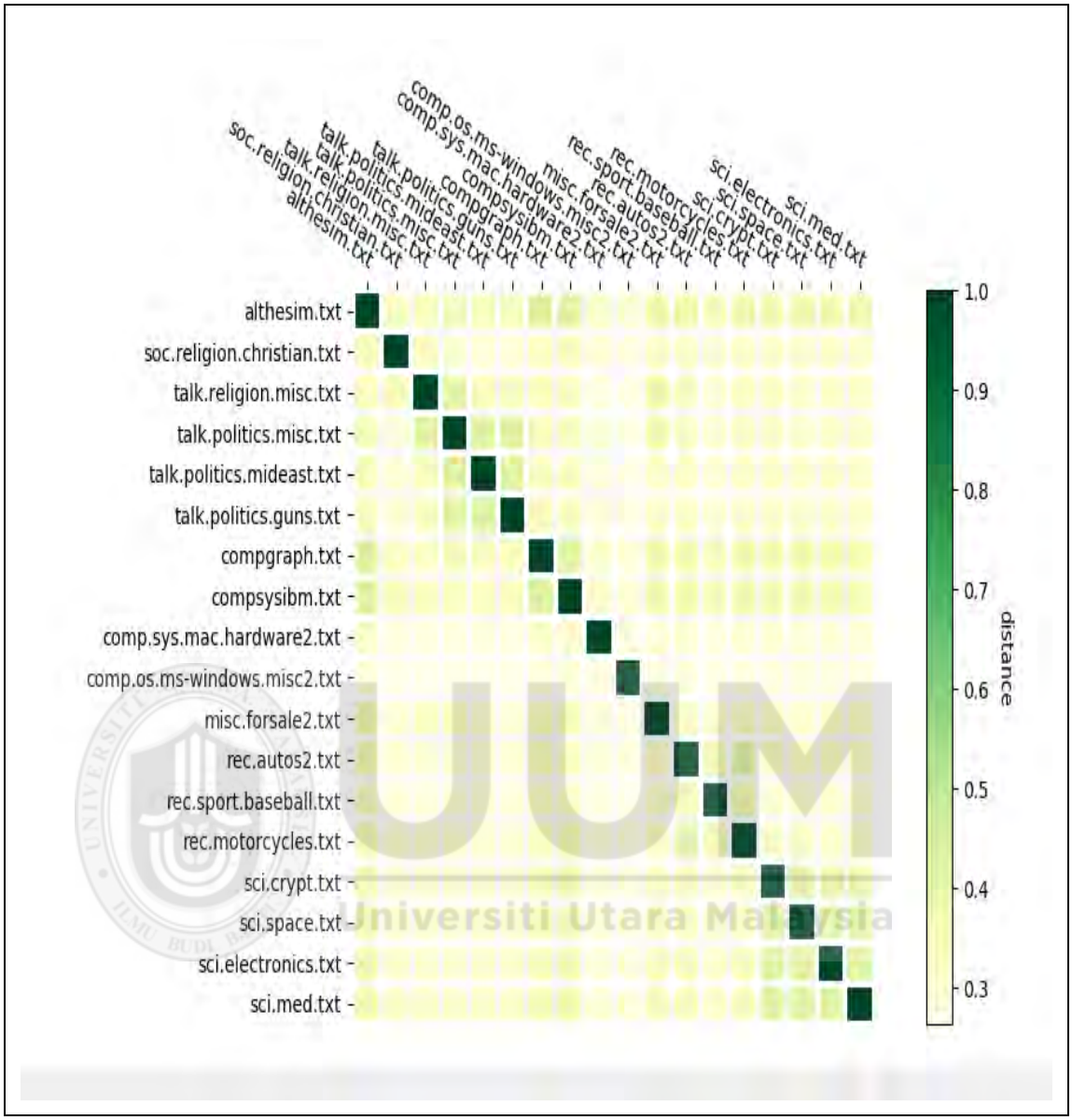

Figure 4. 10. ESET+Manhanttan with marks indicating similar and dissimilar groups

Figure 4.9 and 4.10 shows the results of similar and dissimilar groups of topics using ESET with both Eugene and Manhattan. ESET+cosine seems more relevant compared with other similarity measures in the sense that, it considers the relative frequency of text. It was also noticed that ESET+cosine was able to identify data perfectly. To measure the performance of selected similarity measures with ESET, precision, recall and F1-score of each similarity measures was evaluated manually. These figures were all compared with

a whitepaper by Rennie, (2008) who was able to organize topics into six topics accordingly as shown in (Figure 4.7) from the 20NEWSGROUPS data. In other to evaluate performance of compared measures, red lines were manually drawn as shown in the above diagram to indicate True positive (TP) identified topics while the black line were drawn to show False Positive (FP) topics. ESET+cosine was able to identify 4 similar groups of topics perfectly as shown in Figure 4.8. Other similarity measures were able to identify 3 topics groups correctly. More so, ESET+cosine was also able to identify the most dissimilar topic which is the Misc for sale as was shown in Figure 4.8. Performance evaluation was performed by computing the precision, recall and f-score using Rennie, (2008) as a benchmark study. Table 4.5 shows that ESET + Cosine performed excellently well on text data with topics or focused area of discourse (20newsgroups) than text data (ENRON) with a generalized topic of discussion.

Table 4. 5.

ESET1 results on 20Newsgroups Data

20NEWSGROUPS	Precision	Recall	F1-score
ESET + Eugene Euclidean	0.50	1.00	0.70
ESET + Manhattan	0.50	1.00	0.70
ESET + Cosine	0.92	1.00	0.95

These results undoubtedly showed that readings from cosine is one-sided as shown in Table 4.6.

Table 4. 6.

ESET1 results of 20Newsgroups and ENRON

Data with ESET1	Precision	Recall	F1-score
ENRON ESET+COSINE	0.83	1.00	0.90
20NEWSGROUPS ESET+COSINE	0.92	1.00	0.95
AVERAGE INDEXED SCORE	0.88	1.00	0.93

4.5 Summary

In summary, the generated result proves that optimized cosine is significant and can be further improved to form ESET1. More so, it is imperative that the averaged index F1-score achieved from both data was significantly high but needs to be improved. The ESET1 performs well in identifying text anomalies in documents but however finds it challenging to analyse semantics from documents. This led to the formulation of an optimized cosine measure to consider cosine function as dissimilarity measure which is to detect rare terms in documents and its similarity measures is to detect frequent terms in documents. This was performed because cosine measures best identify similar and dissimilar text on large numbers documents as explained and reviewed in chapter 2&3 (section 2.8 and 3.2) of the study. To further investigate and improve on study results, it is essential to consider improving document feature vectors, term-document matrix and human / corpus based semantic relationship of words for analysing and detecting semantic-based text anomalies present in documents. Thus far, this chapter was able to detect lexical-based or surface level text anomalies more than its semantics although some semantics were identified but more emphasis needs to be made on text content information to model human common-sense knowledge by understanding latent ideas from text documents.

Universiti Utara Malaysia

CHAPTER FIVE

HYBRIDIZING ESET1 WITH LATENT SEMANTIC ANALYSIS FOR SEMANTIC-BASED ANOMALY DETECTION (ESET2)

5.1 Introduction

As regards to this chapter, ESET1 was hybridized with LSA to form ESET2. Before hybridizing ESET1 with LSA model, other topic models were compared and tested for topic/ concept coherence. Topic coherence in this study is aimed at selecting the best model that can be hybridized with ESET1 for semantic analysis and semantic based text anomaly detection in document. Nevertheless, this section lays more emphasis on the implementation of the ESET2 with experimental results and ends with a summary of study findings as will be explained in subsequent sections.

5.2 Comparing models for analysing text semantics

To properly choose a befitting topic model for the ESET2 this study requires an in-depth review about topic models and apparently was a daunting task. Since, understanding trends and concepts from large numbers of text is challenging, it requires complex computation to find relevant concepts and ideas for interoperability and understandability of text documents. Studies have shown that some evaluation methods have improved interpretability of textual models as was discussed in section (3.5). Delving into the quantitative perspective of model evaluation, some researchers made use of *perplexity* or *predictive likelihood* to determine the optimal number of topics to evaluate topic model fit. Perplexity is known to computed by taking the log-likelihood of unseen text documents given the topics defined by a model. A good model has a high likelihood

and resultantly low perplexity. Topic models are evaluated based on their ability to describe documents well (i.e. low perplexity) and to produce topics that carry coherent semantic meaning (Ding, et al. 2016; Drissi & Watkins, 2017). Ding et al., (2018) demonstrated that coherence awareness of topic model shows similar level of perplexity as baseline models but achieves substantially higher topic coherence. Usually, likelihood can be plotted to measure over a spectrum of topics to choose topics that best optimizes measures of choice. On the contrary, perplexity was not computed amongst selected benchmarked models in this study because, perplexity cannot be strongly correlated or sometimes slightly anti-correlated with human judgement (Ramage et al. 2009). Another quantitative solution used in evaluating topic model that has better human interpretability is called *topic coherence*. Topic coherence observes set of words in generated topics and rates the interpretability of these topics. Apparently, there are several measures that calculate topic coherence in various ways. Evidently, studies have stated that topic coherence value proves to be the measure most aligned with human interpretability (Ramage et al. 2009).

In this study, Latent Semantic Analysis, Latent Dirichlet Allocation (LDA) and Hierarchical Dirichlet Processing (HDP) models where incorporated individually with the ESET1 to know which amongst these topic models gives a better performance in identifying coherence in document. As regards to the selected models, it is imperative that HDP is an extension of LDA, both models are designed to address cases where number of mixture components from data (the number of "topics" in document-modelling terms) are not known '*apriori'*. Using LDA for document modelling, all topics are treated as distribution of words in some known vocabulary. For each of these text document, mixture of topics is drawn from a Dirichlet distribution, and then each word in the document is an

independent draw from that mixture. However, the study evaluated topic coherence on 20NG data which was aimed at testing for the most appropriate model that would create text concept consistency, interoperability and will also cope with high dimensional problem in text documents. Nevertheless, coherence was measured on ESET1 using these three models except on Non-negative Matrix Factorization (NMF model) which decomposes multivariate data by creating a defined number of features. However, word interpretability of NMF from existing studies showed that negative values like (height, temperature or even numbers of encounter or visits) are almost nonsensical or simply impossible to compute (Belford et al. 2018; Cai et al. 2008; Giannoulis et al. 2018). Analysing text semantics on topic model with these kinds of final discrete variables may seemingly not perform well especially with the study objectives. However, performance was computed to test for the best suited model (HDP, LDP and LSA) which was hybridized with ESET1 to form ESET2 and their coherence scores are shown in Figure 5.1.

Universiti Utara Malaysia

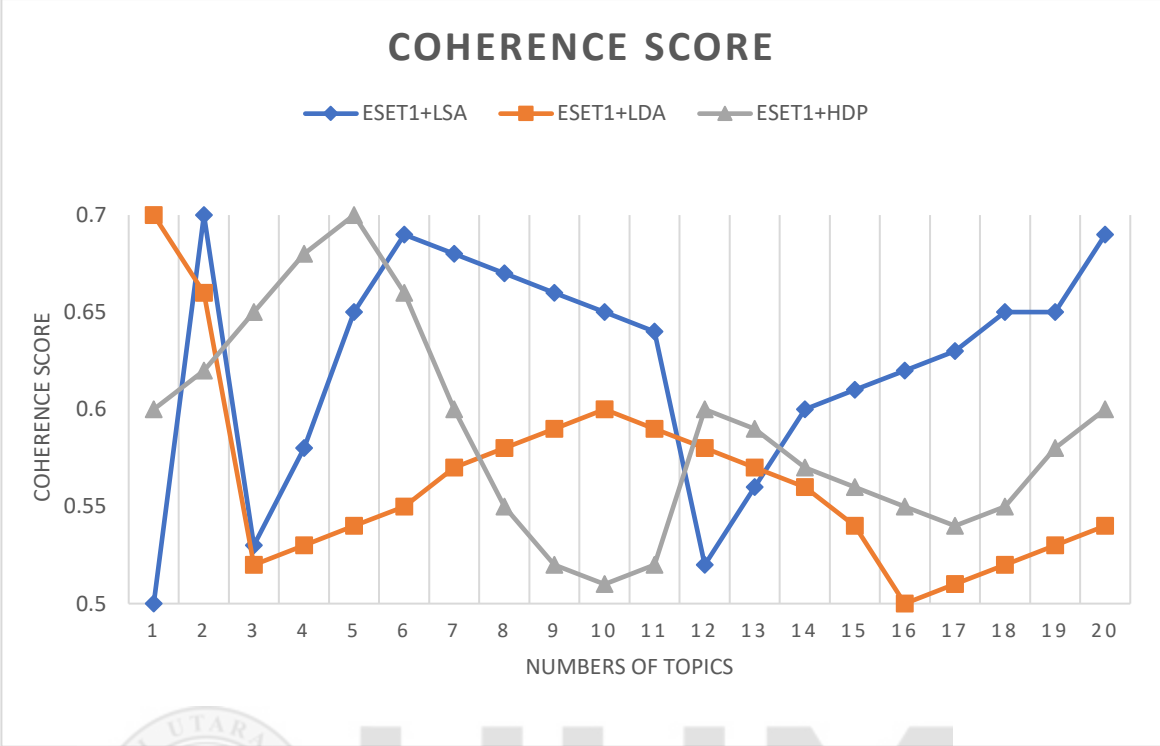

Figure 5.1. Coherence measure for Topic Models

It is pertinent to know that the number of topics for which an average coherence score plateaus, as shown in Figure 5.1 is the spot to be identified. To identify these spots, 20NG data was employed to test coherence score. This is because of the prior knowledge of groups in 20NG data, accuracy can be easily compared and judged by mere looking at coherence graph. Seemingly, ESET1+LSA model was able to score high plateaus from topic 2, 6 and 20 with a decrease in topic 1, 3 and 12. While ESET1+LDA model has its highest plateaus around topic 1 and 2 and ESET 1+ HDP model also exhibited almost similar reading with LDA model by scoring high plateaus around topic 4. Coherence Value (C_V) is found to be most in line with human ratings but can be much slower than U_Mass since it uses a sliding window over the texts. The study made use of C_V and U_mass for evaluating the selected models. Moving on now to choosing numbers of topics

depends on several modalities or factors. ESET1+LSA model has the closest topic coherent match with the identified 20NG data. Many researches prefer to use either LDA or HDP which is because of the high accuracy derived from both models and their probabilistic approach which allows sets of observation to be explained. So far, there are other delimitating factors known with Dirichlet models, which is their inability to capture correlations of topic coherence, also prior knowledge needs to be known over time on fixed number of topics. Additionally, Dirichlet models assumes bag of words to be exchangeable yet sentence structure is not modelled in Dirichlet topics. A number of these limitations have been addressed in papers that followed the original Dirichlet work. Despite its limitations Dirichlet model is central to topic modelling and has really revolutionized the field (Belford et al., 2018; Blei et al., 2012; Boyd-Graber et al. 2007; Chaplot & Salakhutdinov, 2018; Z. Liu, 2013).

This chapter has demonstrated that employing topic coherence is useful in evaluating how models identifies consistency in semantic correlation of text documents. To identify semantics and anomalous data, coherence level of semantics in documents needs to be well-analysed without falling into issues relating to having prior knowledge about text data. LSA model with ESET1 according to the obtained coherence scores has shown the closest coherence readings with the actual 20NG data therefore, ESET1+LSA is leveraged in this study to detect semantic based text anomaly.

5.3 Hybridizing ESET1 with Latent Semantic Analysis (LSA)

This phase is poised at hybridizing the modified ESET1 with LSA to form ESET2. LSA was used to overcome the problems of analysing semantics in traditional lexical matching (Zhang et al., 2016). Consequently, LSA is usually criticized as with having a low

discriminative power for representing text documents although it has been validated with good representative quality. The study made use of Singular Value Decomposition (SVD) as described in section (2.9) to improve the discriminative power of ESET2. More so, SVD was applied on terms which theoretically explains that the dimension expansion of document vectors and dimension projection are manipulations of terms. Figure 5.2 describes the steps involved in creating the ESET2.

Step 1: Pre-processing Text
- Import text data from repository and prepare text by cleaning, tokenizing
- Resultant text devoid of stop-words were displayed and fed as a list.
- Make resultant text into bigrams and lamentize text.

Step 2: Detecting similar and dissimilar textual data
- Resultant terms were vectorized to create a term matrix
- Term-document similarity and dissimilarity was identified using cosine function

Step 3: Models were created from documents to analyse text semantics
- Best model was chosen, and visualizations of terms were performed on identified model to evaluate term coherence dominant and rare documents were identified.

Figure 5.2. Steps in ESET2

Universiti Utara Malaysia

Distribution of word was performed to check for relevant terms as well consider smoothing or pruning of terms for the sole purpose of tuning parameters of the technique to get better accuracy as shown in Figure 5.3

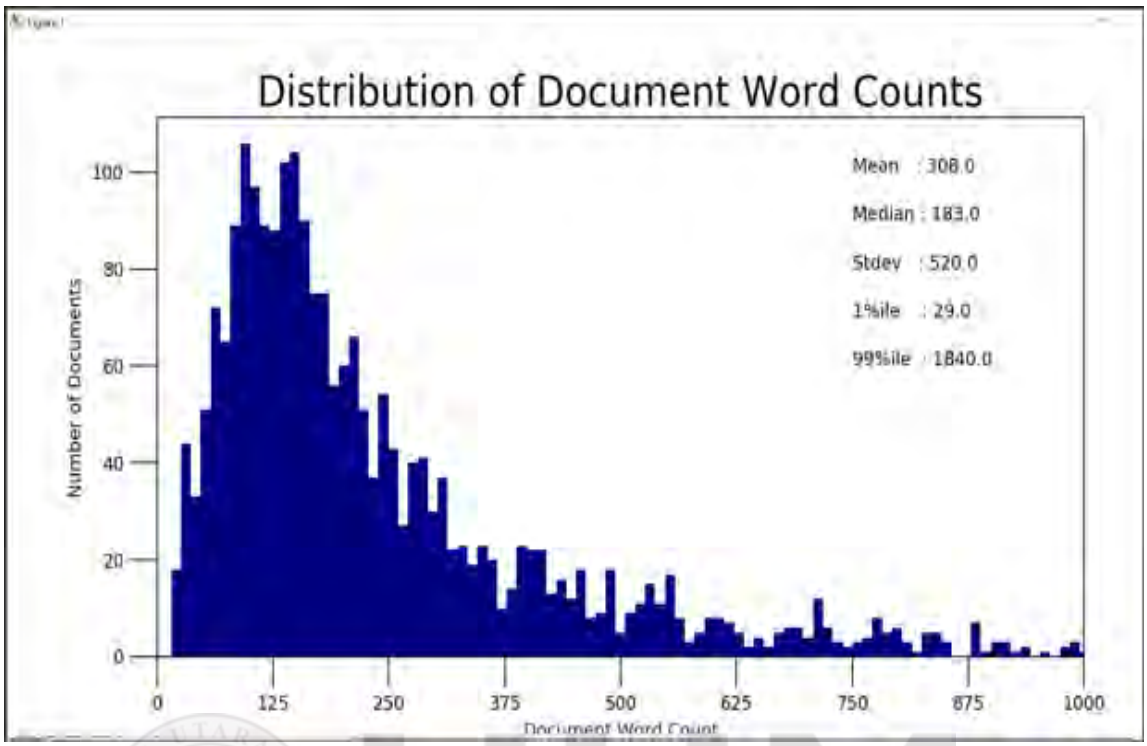

Figure 5.3. Distribution of Documents word counts using the 20NGs data

It is key to know the distribution of word counts when dealing with documents with large numbers of text. It gives a better understanding of identifying samples of terms that are most represented in documents. Afterwards, ESET2 is applied to identify terms that are semantically related in documents. This is achieved by initially computing text similarity as shown in Table 5.1.

Table 5.1.

Similarity Score of data

No	Data	Similarity Score
1	20NG	[[1. 0.02373208 0.03562947 0.01365786 0.00745555 0.00647735]]
2	ENRON	[[1. 0.65259454 0.66396438 0.77105269 0.5575983 0.86160613 0]]
3	NIPS	[[1. 0.65259454 0.66396438 0.77105269 0.5575983 0.86160613]]

Table 5.1. shows the similarity between documents. ESET2 made use of SVD on data to solve issues relating to low rank matrix approximation from the approximated term-

document matrices. Three steps were involved in invoking SVD in ESET2 as shown in Figure 5.4.

Step 1: Given C, construct its SVD in the form $C=U\Sigma V^T$.

Step 2: Derive from Σ the matrix Σ_k formed by replacing by zeros the $r-k$ smallest singular values on the diagonal of Σ.

Step 3: Compute and output $C_k = U\Sigma_k V^T$ as the rank-kk approximation to C

Figure 5. 4. Distribution of Documents word counts using the 20NGs data

Where C is the term-document matrix and U, Σ and V^T are SVD computed matrices. The unseen terms derived by SVD from documents can be transformed using genism to plot similarity and dissimilarity functions present in ESET2 to detect data with close or far semantic relationship based on Term- document matrices. The technique leveraged SVD to display words. A 2D graph of documents and words in Figure 5.5, 5.6 and 5.7 shows similar and dissimilar data using words and documents as a representative of ESET2 leveraging SVD with topic components to display similar terms.

Universiti Utara Malaysia

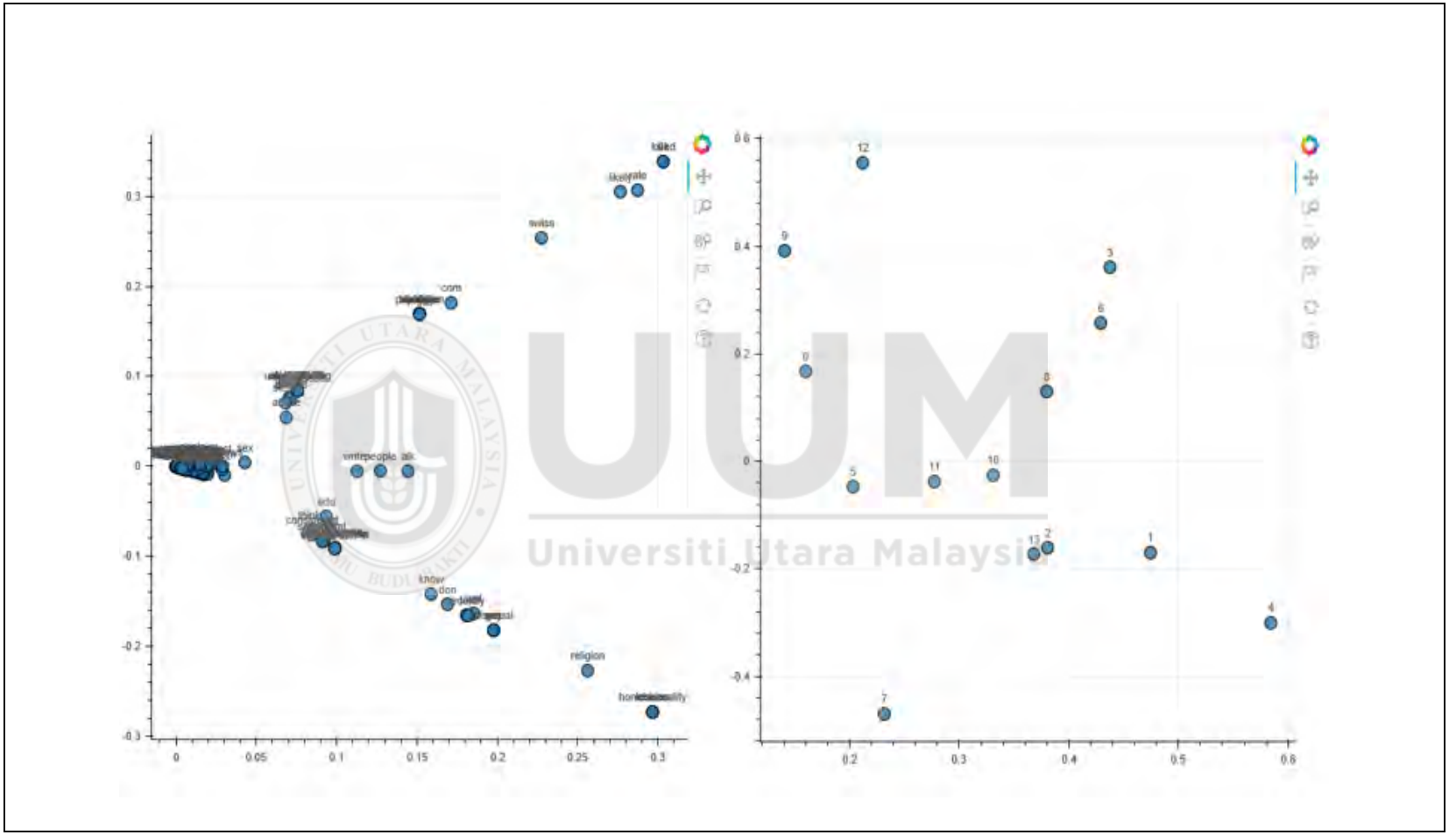

Figure 5. 5. Distribution of terms in 20NGS data

Universiti Utara Malaysia

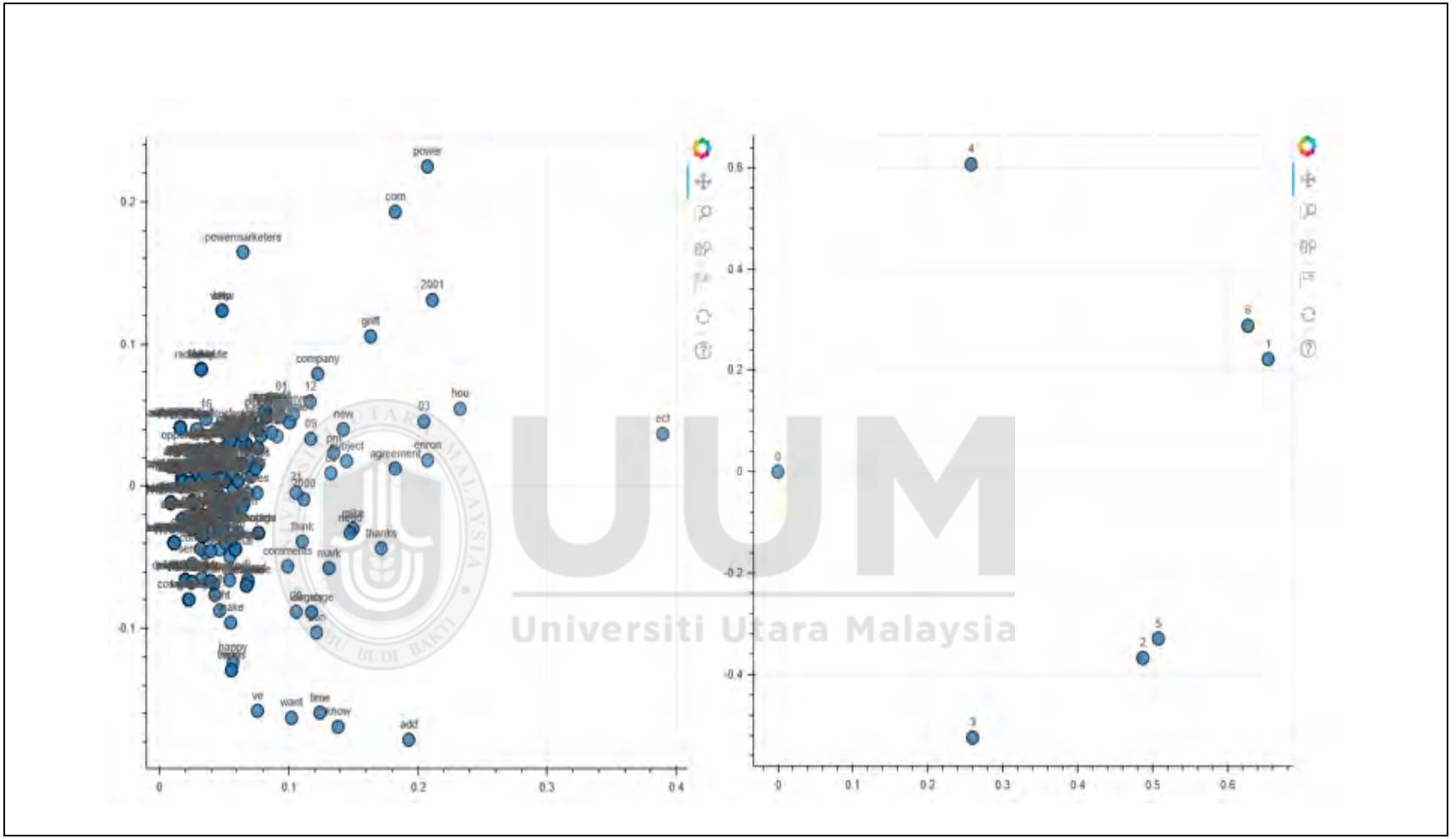

Figure 5. 6. Term distribution of ENRON mail messages

Universiti Utara Malaysia

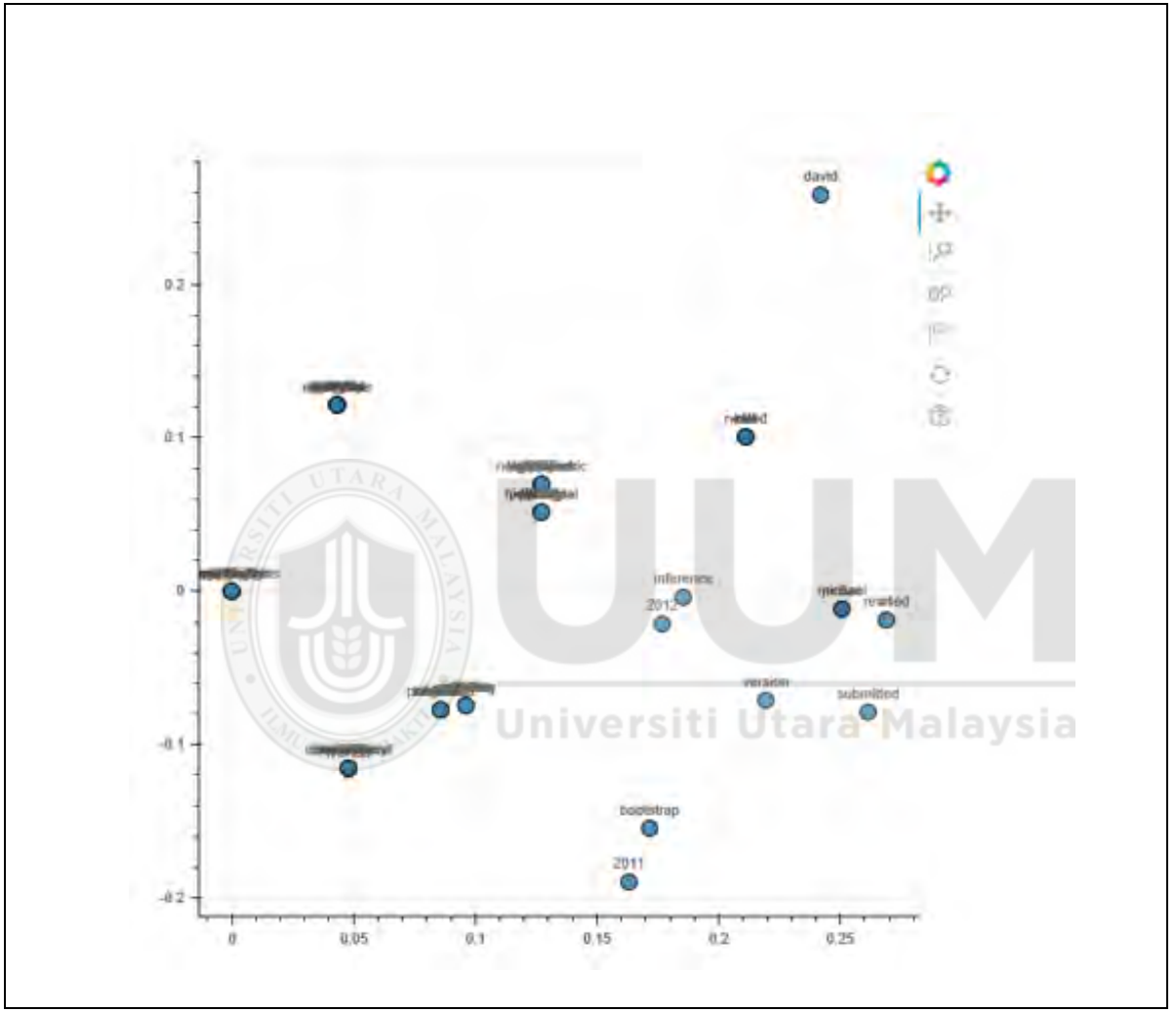

Figure 5. 7. Term Distribution of NIPS data

Universiti Utara Malaysia

To generate good word representation, a significantly larger number of documents are needed. The generated data were grouped into 6 concepts and in every concept contained 10 words that shows almost all necessary information about data. For instance, it was noticed that ENRON is a power organization engaging it staff members in series of meetings, making legal agreements that is based on time and accounting terms and conditions. NIPS on the other hand, have editions or versions of revised articles by different authors concentrating on how massive useful data can be computationally modelled for use. Lastly, is the 20NG data, As regards to this study, 20NGs data was mixed with an outlying data (Agro-food terms glossary) which carries an entirely different information from the original 20NG data. Detecting anomalous data in this phase was performed by running the ESET2 at a maximum difference of 0.02 and a minimum difference of 0.05 for 5 different iterations. It was noticed that almost all agro-terms were identified. Therefore, ESET2 was noticed to achieve high precision, recall and F-measure.

Universiti Utara Malaysia

After examining the false positive score from the mixed 20NGS data, it was noticed that Agro-food term glossary data were different from groups of topics in 20NGS. This indicated that the identified Agro-terms were termed as anomalous from the original 20NGS data. More importantly, word representative in documents generated by ESET2 showed an interesting observation as shown in Table 5.2.

Table 5. 2.

List of recognized concepts in ESET +LSA

ENRON Data					
Concept 1	**Concept 2**	**Concept 3**	**Concept 4**	**Concept 5**	**Concept 6**
Ect	Enron	Com	Believe	griff	20
hou	james	Power	committee	company	language
2001	message	marketers	draft	gray	rusty
enron	original	http	revised	mary	need
power	award	www	mark	master	consumers
03	thank	know	martin	customer	add
add	list	enroncredit	mike	received	agreement
NIPS Data					
Revised	Acceleration	30	19	2014	bootstrap
v1	achieving	answers	apr	25	21
submitted	ccelerated	based	ayesian	chong	28
jordan	central	build	charles	develop	ameet
michael	extensions	compositional	interne	dirichlet	ariel
david	generalizations	dan	internet	nested	data
version	gradient	dependency	jan	hierachy	dec
20NGS Data					
Killed	Pc	Atheist	Navy	baseball	Fruits
choice	windows	fish	rules	apple	Animal
homosexu	edu	ripem	presentations	regions	husbandry
ality	comp	atheism	scientific	march	local padi
lesbian	ibm	darwin	seminar	admiral	rice
rate	mouse	resources	visualization	baltimore	plant
likely	sys	navy	reality	blue	palm oil

Table 5.2. Shows the concept generated from ESET2 using NIPS, ENRON text

5.4 Performance evaluation of ESET2 with results

Performance evaluation shows that the total number of correctly identified anomalous terms as true positive and the total number of incorrectly identified number of anomalous terms as false positive. The range of values is from (0-1), where 1 is the optimal value used in evaluating the harmonic mean score (f-measure or score). More importantly, the evaluation metrics described in this study assess topic model quality with their theoretic

methods to find best topic models. However, the best method to assess quality topic model practically is the Human-in the-loop approach, where an expert must manually spot topic intruders in documents. ESET2 achieved a precision of 0.88 or 88% with an overall accuracy score of 0.93 as shown in Table 5.3.

Table 5.3.

Evaluation Metrices of ESET2 on 20NGS

ESET + LSA	**Evaluation Metrices**
Precision	0.88
Recall	1.00
F-measures	0.93

Nevertheless, the ESET2 performance metrics computed the homogeneity, completeness and v-measures. These metrices were aimed at computing the accuracy of semantic based text anomaly detection in document. Homogeneity represents the objective that each concepts or topics from documents contains only members of a ground truth group and Completeness represents the objective that all members of a ground truth group are assigned to the same concept.

Universiti Utara Malaysia

The higher the homogeneity score the better the results, in this case the higher the homogeneity the lesser heterogeneous data. V-measure is an entropy-based measure which explicitly measures how successful the criteria of homogeneity and completeness have been satisfied which is equivalent to Normalized Mutaual Information (NMI). This has helped in providing an elegant solution to accurately evaluate and match terms into data points as it was identified by ESET2. Results of ESET2 were presented and averagely scored to distinctly show the performance of all metrics used as was shown in Table 5.4.

Table 5.4.

ESET2 Evaluation Metrices for 20NEWSGROUPS data

Evaluation Metrices	ESET2 results
Homogeneity	0.92
Completeness	0.82
V-measures	0.86
Adjusted Rand Index	0.52

ESET2 was benchmarked with GSDPMM (Yin & Wang, 2016) using the 20NGS data as will be shown in Table 5.5. This means that the study comparison in this phase is only based on long text document whereby short term documents were completely ignored. It is important to know that measures such as V-measure and Adjusted Rand Index were based on information theoretic evaluation and hence cannot be affected by the curse of dimensionality present in text data.

Table 5.5.

ESET2 benchmark results

Techniques	Homogeneity	Completeness	V-measure
GSDPMM. Yin & Wang, (2016)	0.4	0.7	0.51
ESET2	0.92	0.82	0.86

Universiti Utara Malaysia

5.5 Summary

Thus far, this chapter argued that the performance of ESET2 has proven to be successful in identifying semantic similarity of terms, but it is yet to perform as expected in detecting outlying or anomalous text data as compared with GSDPMM which performed better with an accuracy or F-measure of about 96% while ESET2 reads a lower accuracy of 93%. To further improve on ESET2 results, it is necessary to consider optimizing document feature vectors, term-document matrix and human / corpus based semantic relationship of words.

To conclude this section, improvement should be made on detecting semantic based text anomalies by identifying semantical and syntactical relationships of words in both short and long sentences or documents. This can be achieved by employing both the knowledge and corpus-based approach as will be further discussed in the next chapter of the study.

ESET2 employs a linear model, which may not perform well on data with non-linear dependencies. It assumes a gaussian distribution of terms in documents which may not be relatively true for all problems such as identifying intrinsic semantic similarities and dissimilarities in terms. Also, ESET2 employs SVD to identify semantics which has its own major weakness especially when trying to disambiguate senses in document. Until a better solution is found for identifying intrinsic semantic dissimilarity/similarity in words, then a better performance can be achieved. A need for a better word sense disambiguation is desired for semantic-based text anomaly detection for better performance.

Universiti Utara Malaysia

CHAPTER SIX

INTEGRATING WSD ALGORITHMS WITH ESET2 (ESET3)

6.1 Introduction

This study follows series of steps in proving that rare and frequent terms in documents are considered anomalous depending on their contextual usage. This was proven by developing ESET technique. ESET consist of 4 major phases that are aimed at detecting semantic-based text anomaly in large numbers of document as was illustrated and described in section (3.2) of this study. In the previous chapter, ESET2 was unable to efficiently tackle issues relating to synonyms and polysemous words as envisaged from the study objectives.

It is essential to consider text canonization, which has proven to provide better solution in resolving issues related to text ambiguity as reviewed in section (2.10). The study proposes a combined Word Sense Disambiguation algorithm that was integrated with ESET2 to canonize text for semantic disambiguation as well detect semantic-based text anomalies in documents. Apparently, word disambiguation approaches are classified into structure, knowledge and corpus-based approaches. It is vital to have a background knowledge as to why the study combined all approaches to form ESET3.

6.2 Integrating combined WSD algorithms with ESET2

The combined WSD algorithms namely Lesk and Selectional preference considers text as sequence of words by dealing separately with words in sentences according to their

semantic and syntactic structure. The information content of word is somehow related to the frequency of the meaning of word in a lexical database or corpus. Ambiguated terms or sense were checked in this study by adopting Lesk algorithm which makes use of the dictionary definition to find synonymous terms (synset). This approach helps to relate words with their closest meaning using dictionary definitions. It is a corpus based approach and has been used by many studies to resolve polysemy and synonym related problems in text documents (A.Rajaraman, J. Leskovec, 2016; Basile et al., 2014; Lesk, 1986; Sardar, 2018).

More so, this study adopts Selectional Preference (SP) algorithm as a knowledge-based approach of word disambiguation. This algorithm makes use of Part of Speech (POS) tags like Nouns, Proper Noun, Possessive words, Adverb and Verbs frames in wordnet and agent- action -object of words to disambiguate senses inn documents. It requires an exhaustive enumeration of argument structure of verbs to infer meanings to words that does not have dictionary meaning, pronoun resolution, named entity recognition, semantic role labelling and information extraction. Both adopted WSD algorithms were combined to tackle issues relating to text ambiguity.

Consequently, the combined WSD algorithms follows a structural hierarchical and non-hierarchical relatedness of words to tackle semantic related issues. Further explanation can be seen from flowchart in Figure 6.1 which illustrates steps and functions of every necessary process and phases involved in disambiguating text from the WSD algorithm.

Universiti Utara Malaysia

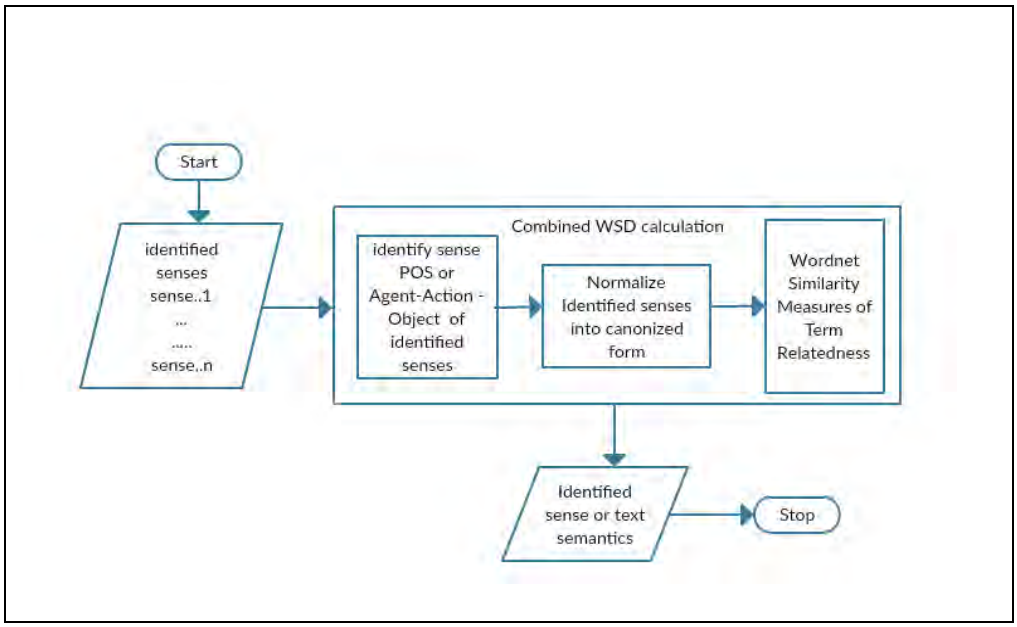

Figure 6.1. Combined WSD flowchart

This technique in Figure 6.1 is implemented following the steps outlined in Figure 6.2

Step 1: Text were extracted from repository (bag of words) and NLP was applied to stem, and normalize text data
Step 2: Resultant text devoid of stop-words were displayed and fed as a list.
Step 3: Document term matrix was constructed using an optimized cosine similarity function in ESET1 using LSA
Step 4: Acquired terms were modelled into topic/ discourse and Terms were canonized by combining two WSD algorithms to resolve issues relating text ambiguities.
- Part of speech were taking into consideration such as the Noun, Verb, Determinants and possessive words are considered,
- Subjectivity of words namely action, object rule of words is also considered to disambiguate words

Steps 5: Finally, actual terms with synsets from the combined WSD algorithms were evaluated using standard performance metrics

Figure 6.2. Combined WSD steps

With the increasing research in sentence comparison, many data have been built to test for the performance of sentence semantic similarity and dissimilarity. According to Li, et al (2009) compared sentence similarity based on the information gotten from sentence such as the, Objects-Specified Similarity (OSS), Objects-Property Similarity (OPS),

Objects-Behaviour Similarity (OBS) and Overall Similarity (OS) as was explained in section (2.9). These aspects of similarity are defined to determine sentence semantic similarity so as to have the information of knowing deviating or anomalous sentences from documents as reviewed in (Li, et al 2009).

An experiment was made with randomly selected sentence samples (four positive and four negative paired sentences) with their corresponding similarities scores that was obtained from MS paraphrase test corpus (Li, et al 2009). The paraphrased test corpus was compared with ESET3 to demonstrate the feasibility and validity of how sense disambiguation works in detecting semantic based text anomalies in short sentences is as shown in Table 6.1.

Universiti Utara Malaysia

Table 6. 1.

Sample sentence for semantic similarity

No	Sentences
	Pair of similar sentences
1	*"Taha is married to former Iraqi oil minister Amir Muhammed Rasheed, who surrendered to U.S. forces on April 28." "Taha's husband, former oil minister Amer Mohammed Rashid, surrendered to U.S. forces on April 28."*
2	*"On July 22, Moore announced he would appeal the case directly to the U.S. Supreme Court." "Moore of Alabama says he will appeal his case to the nation's highest court."*
3	*"Six Democrats are vying to succeed Jacques and have qualified for the Feb. 3 primary ballot." "Six Democrats and two Republicans are running for her seat and have*
4	*qualified for the Feb. 3 primary ballot."*
	"Agriculture Secretary Luis Lorenzo told Reuters there was no damage to the vital rice crop as harvesting had just finished." "Agriculture Secretary Luis Lorenzo said there was no damage to the vital rice crop as the harvest had ended."
	Pairs of dissimilar sentences
5	*"A soldier was killed Monday and another wounded when their convoy was ambushed in northern Iraq." "On Sunday, a U.S. soldier was killed and another injured when a munitions dump, they were guarding exploded in southern Iraq."*
6	*"Perkins will travel to Lawrence today and meet with Kansas Chancellor Robert Hemenway." "Perkins and Kansas Chancellor Robert Hemenway declined comment Sunday night."*
7	*"'I am proud that I stood against Richard Nixon, not with him,' Kerry said." "'I marched in the streets against Richard Nixon and the Vietnam War,' she said."*
8	*"The report by the independent expert committee aims to dissipate any suspicion about the Hong Kong government's handling of the SARS crisis." "A long-awaited report on the Hong Kong government's handling of the SARS outbreak has been released."*

Four aspects as earlier mentioned are defined to determine sentence similarities. First, two compared sentences are respectively chunked with noun phrases and verb phrases.

Secondly, for each sentence, all nouns in noun phrases are chosen as the objects specified in the sentence, all adjectives and adverbs in noun phrases are chosen as objects properties and all verb phrases are chosen as the object's behaviours. Then, the four similarities are calculated based on a semantic vector method.

The result of this can further be illustrated by presenting sentence from the compared semantic similarity measures with the ESET3 as will be shown in Table 6.2.

Table 6. 2.

Comparison of semantic Similarities results with ESET3

Sample Pair	Original decision with similarity score thresholds	OS	OP	OB	Overall	ESET3
Sample1	(0.5-1.0) similar	0.87	1.0	0.58	0.85	0.93
Sample2	(0.5-1.0) similar	0.85	0.0	0.99	0.58	0.91
Sample3	(0.5-1.0) similar	0.97	1.0	0.71	0.93	0.90
Sample4	(0.5-1.0) similar	1.0	1.0	0.87	0.98	0.98
Sample5	(0.0-0.4) Dissimilar	0.58	0.0	0.5	0.38	0.47
Sample6	(0.0-0.4) Dissimilar	0.72	0.0	0.2	0.38	0.20
Sample7	(0.0-0.40) Dissimilar	0.50	0.0	0.8	0.38	0.31
Sample8	(0.0-0.4) Dissimilar	0.88	0.0	0.0	0.39	0.10

Universiti Utara Malaysia

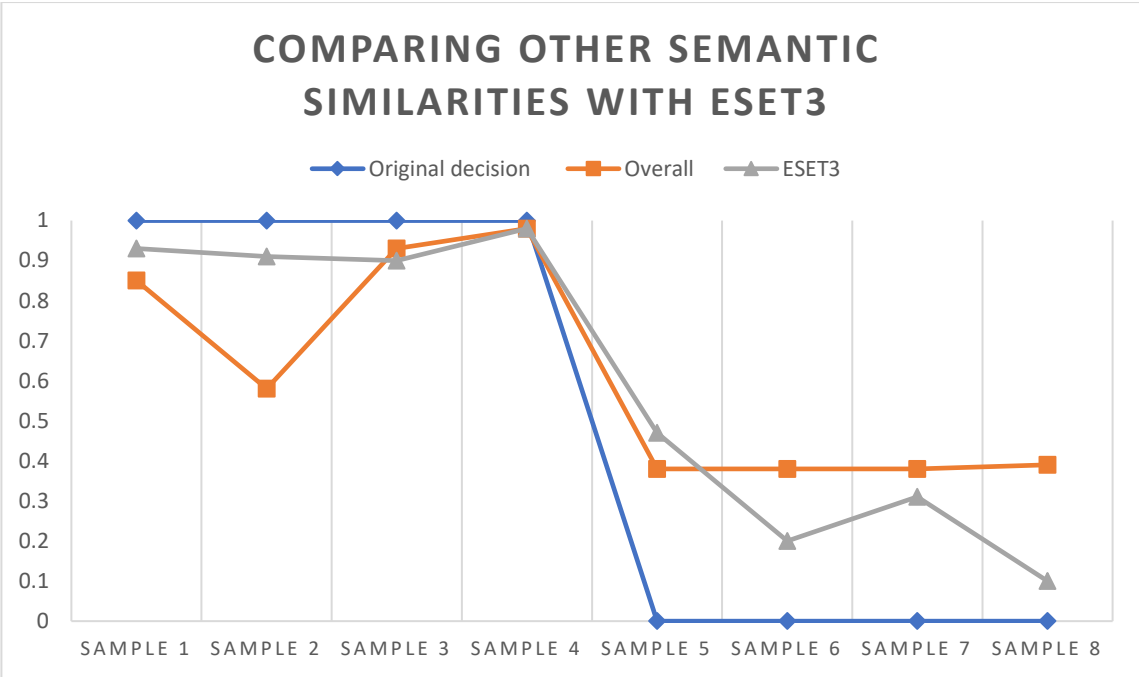

Figure 6.3. Results of compared similarity measures with ESET3

Table 6.2 and Figure 6.3 Shows that word pairs similarity scores from 0.5 -1.0 are relatively similar and word pairs from 0.0-0.4 are relatively dissimilar. Experiments shows that ESET3 identifies sentence semantic similarity comparison more intuitive and render a more reasonable result, which imitates human comprehension to the sentence semantics in documents (Li, et al 2009). To further test ESET3, detection of semantic based text anomalies in document was performed as will be discussed in the next sub-section.

Universiti Utara Malaysia

6.3 Performance evaluation of ESET3 with results

Another experiment was conducted on ENRON, NIPS KOS data. Detecting the occurrence of anomalous text that would help infer a shift in behavioural pattern also known as side information is aimed to be achieved in this section. ESET3 is benchmarked with an interesting study made by Mahapatra et al. (2012), who used a novel context-detection algorithm that operates on an external corpus of general semantic relationships

allowing the identification of words and topics occurrences (Wordnet and NGD) to an LDA-based text modelling approach for anomaly detection.

Before comparing ESET3 with (Mahapatra et al., 2012), it is imperative to know that the technique used by Mahapatra et al., (2012) was based on detecting data that was made up of user generated exemplars. This shows that the idea could be implemented at various level of abstraction (Topic and word level). Additionally, (Mahapatra et al., 2012) made use of an augmented dummy topic in the contextual database to detect previously undetected anomalies based on selective threshold manipulations of operators input. This study also based its detection on the identified normal and atypical topics identified by Mahapatra et al., (2012) using same exemplary generated data to detect semantic- based text anomalous data present in documents.

This section aims to detect previously undetected semantic-based anomalous textual information and to reduce false positivity during performance evaluation. The objective is to incorporate real-world semantics to textual data-streams and logs. Threshold parameter for empirical tuning of semantic-based text anomalies are based on the tagged topics identified by some domain experts as described in (Mahapatra et al., 2012). However, ESET3 made use of both the hierarchical and non-hierarchical word relationship leveraging the Latent semantic word to vector models and WSD algorithms to disambiguate senses. Results was actualized using the precision, recall and f1-score. The benchmarked ESET3 with Mahapatra et al., (2012) were tested on four data and snippet results are shown in Table 6.3.

Table 6. 3.

Snippet of some generated semantic based text anomalies detected.

Flagged atypical topics by Human Experts	ESET3 identified terms with similarity index		
Mixed 20NGS with agro-food terms			
Farming, animal husbandry and livestock	{('livestock', 'fruits'): 0.07692307692307693, ('farming', 'fruits'): 0.09090909090909091, ('livestock', 'rice'): 0.08333333333333333, ('farming', 'rice'): 0.07692307692307693, ('livestock', 'plant'): 0.125, ('farming', 'plant'): 0.1, ('livestock', 'palm'): 0.08333333333333333, ('farming', 'palm'): 0.08333333333333333, ('livestock', 'oil'): 0.06666666666666667, ('farming', 'oil'): 0.08333333333333333, ('livestock', 'livestock'): 1.0, ('farming', 'livestock'): 0.06666666666666667, ('livestock', 'cocoa'): 0.0625, ('farming', 'cocoa'): 0.08333333333333333, ('livestock', 'fruit'): 0.07692307692307693, ('farming', 'fruit'): 0.09090909090909091}	[('fruits', 'NNS'), ('animal', 'JJ'), ('husbandry', 'JJ'), ('beverage', 'NN'), ('local', 'JJ'), ('padi', 'NN'), ('rice', 'NN'), ('plant', 'NN'), ('palm', 'NN'), ('oil', 'NN'), ('livestock', 'NN'), ('cocoa', 'NN'), ('fruit', 'NN')]	Similarity index value : 0.55 Similar
NIPS			
Processing speech recognition	{(‘speech’, ‘speech’): 1.0, (‘speech’, ‘context’): 0.1111111111111111, (‘speech’, ‘processing’): 0.1111111111111111, (‘speech’, ‘data’): 0.16666666666666666, (‘speech’, ‘function’): 0.16666666666666666, (‘speech’, ‘words’): 0.5, (‘speech’, ‘speaker’): 0.09090909090909091}	[(‘speech’, ‘NN’), (‘context’, ‘NN’), (‘recognition’, ‘NN’), (‘processing’, ‘VBG’), (‘data’, ‘NNS’), (‘function’, ‘NN’), (‘words’, ‘NNS’), (‘speaker’, ‘NN’)]	Similarity index value : 1.0 Similar
KOS			
Iraq war	{('state', 'soldiers'): 0.08333333333333333, ('war', 'soldiers'): 0.07142857142857142, ('iraq', 'soldiers'): 0.0625, ('state', 'war'): 0.3333333333333333, ('war', 'war'): 1.0, ('iraq', 'war'): 0.06666666666666667, ('state', 'weapons'): 0.1111111111111111, ('war', 'weapons'): 0.14285714285714285, ('iraq', 'weapons'): 0.07142857142857142}	[('soldiers', 'NNS'), ('iraq', 'VBP'), ('war', 'NN'), ('defense', 'NN'), ('united', 'VBD'), ('military', 'JJ'), ('weapons', 'NNS'), ('abu', 'VBP')]	Similarity index value : 0.50 Similar

Universiti Utara Malaysia

As shown in Table 6.3 Semantic based anomalous texts were identified from different topics in ENRON, NIPS and KOS data. This study made reference to the human experts evaluation in Mahapatra et al., (2012) whom were able to flag 10 topics as semantically anomalous in ENRON data namely; legal agreement, spam mail messages, system maintenance, discussion amongst new MBA graduates and others. ESET3 was able to identify 7 topics out of 10 topics identified by Human experts.

This was due to human names present in some diverting topics which was not easily detected or captured. ESET3 was unable to disambiguate names or find related meaning to attach human name. Identifying and relating polysemy and synonymous words were focused more on other terms this made the false positive ratio of detection accuracy higher than expected. Name Entity Recognition (NER) would have performed perfectly well in this case. ESET3 on ENRON obtained a significant score of 0.82 accuracy. Moving on to the 20NG data, it was easy to detect anomalous topics if compared with the ENRON data. This was apparently due to the groups of distinct topics available in the data. 20NGs was manually mixed with an outlying document (Agro-food term glossary).

Universiti Utara Malaysia

The Agro-food term glossary was tagged anomalous in 20NG and the ESET3 was able to identify all terms related to this topic reducing false positive to almost zero. NIPS data on the other hand, was tasking compared to other data. This was due to presence of special keywords and the presence of almost unnoticeable or little anomalous topics. 5 topics were flagged anomalous by human experts. The ESET3 was able to detect 3 anomalous topics. Accuracy was not as high as compared with 20NGS or ENRON data, scoring 0.75 accuracy. Lastly, Daily KOS blogs had more anomalous or diverging topics detected because terms were not specific to domain discourse compared to NIPS data. This made

it easy for the ESET3 to identify flagged topics namely Iraq war, Gay rights, George Bush fight against Terrorism, Abu Ghairub incident in prison and discussion about the White House. However, terms from these topics are closely related and as well share almost a common theme which is mainly discourse about American Politics. The accuracy of the ESET3 is measured using precision recall and F-1 Measure. ESET3 was performed under many conditions which mostly involve tuning of parameters and as well manual selection of some data like in the case of 20NGS. ESET3 was able to provide better accuracy score when compared with ESET2. Leveraging the 20NGS data for both techniques, ESET3 obtained a highly significant score of 0.97 accuracy in comparison to both ESET2 and GSDPMM where 0.93 and 0.96 accuracy was obtained respectively.

It was noticed that ESET3 outperformed both ESET2 and GSDPMM techniques with an accuracy score of 0.97. This has indicated that ESET3 has viable potentials in detecting semantic-based text anomaly. Although some adjustments need to be made to fine tune parameters in ESET3 for the purpose of improving its performance as shown in Table 6.4.

Universiti Utara Malaysia

Table 6. 4.

ESET3 benchmark experimental results

Techniques		**Data**	**Precision**	**Recall**	**F-1 Score**
Mahaputra	1	ENRON with WORDNET	0.89	0.80	0.84
	2	NIPS with WORDNET	1.00	0.8	0.88
	3	KOS with WORDNET	0.80	0.57	0.67
ESET3	1	ENRON	0.82	0.85	0.82

	2	NIPS	0.75	0.72	0.75
	3	KOS	0.98	0.90	0.93

6.4 Enhanced Exception Technique (ESET3)

ESET is heavily focused on text data, semantic based anomaly detection and representation of text data. This was performed by integrating and hybridizing existing models, algorithms and techniques to achieve the study objectives. ESET is mainly aimed at knowledge creation, sharing as well as decision making purpose from text data. In comparison with SET, some distinguishable operations were noticed. Table 6.5 shows the comparison of SET and ESET

Table 6. 5.

Comparing SET with ESET

S/N	SET	ESET
1	Used on categorical data	It can be used on both text and categorical data
2	Detect anomaly in log files found in large database	Is employed on text corpus
3	Cardinality counts only	Cardinality performs count vectorization on textual data which enables pruning documents or text frequency.
4	Dissimilarity functions made use of variance and standard deviation.	Dissimilarity functions with Optimized Cosine dissimilarity/similarity.
5	Used for detecting deviants in log files	It is used for detecting semantic-based text anomaly
6	Consider only dissimilar or deviating data as anomaly	Considers both similar and dissimilar data as anomaly depending on the context of usage.

The comparison between SET and ESET shows that ESET has a different study objective.

Universiti Utara Malaysia

Before conclusions are made, it is also important to highlight how ESET benchmarked studies to satisfy the research objectives mentioned in (section 1.3). Table 6.6. shows the benchmarked studies with the research objectives.

Table 6. 6.

ESET benchmark experimental setup

ESET	Benchmarked studies	Research Objectives
ESET1	Suspicious departments in ENRON by (Gloor et al.,2006), (Godbole, 2002) and Identifying groups in 20NGs data by (Rennie, 2008)	Test for dissimilarity and similarity of textual data in corpus RO2 and RO5
ESET2	A model-based approach for text clustering with outlier detection (abbr. to GSDPMM). (Yin & Wang, 2016)	Detect semantic-based text anomaly RO3
ESET3	Identifying Sentence semantic similarity by (L. Li, Hu, Hu, Wang, & Zhou, 2009) and Detecting contextual based text anomaly(Mahapatra et al., 2012)	To identify canonical forms in detecting semantic-based text anomaly in documents? ESET RO4

Universiti Utara Malaysia

6.6 Summary

As was pointed out in this chapter, the study was benchmarked against study performed by Mahaputra using almost same data and evaluation measures was aimed at improving detection accuracy of semantic based text anomalies in documents. This study made it clear that ESET is completely different from SET in terms of objectives and techniques. It was also noticed that the integration of the combined WSD algorithms used in this study has improved disambiguation and analysis of text semantics perfectly as experimented on with the sample data from MS paraphrase test corpus. Recently, researchers are beginning to show interest in detecting fake news, identifying filter bubbles as well using an unsupervised approach to swiftly detect semantic based text anomalies in documents.

However, detection of semantic based text anomalies may seem tasking for huge numbers of documents. A representation scheme is needed to understand and map out knowledge from documents. To put ESET to test, a representation scheme was leveraged in the next chapter for easy interpretation of information based on the detected semantic based text anomalous datasets identified in this study.

Universiti Utara Malaysia

CHAPTER SEVEN

REPRESENTATION SCHEME FOR THE IDENTIFIED SEMANTIC-BASED TEXT ANOMALIES

7.1 Introduction

As far as ESET3 is concerned, understandability and interpretation go a long way with the achieved results for knowledge sharing purposes. Unless ways are identified to increase context to state-of the-arts techniques, there will always be reluctance of researchers not adopting these techniques. In front of such restraints, researchers are more focused on the development of interpretable representation schemes.

7.2 Representation scheme for ENRON Data

Visualizing Semantic-based text anomalous data from ESET3 follows a basic concept according to the benchmarked study. In this chapter, Concept Network Graph with FOL were adopted and used to represent identified anomalous topics. The representation scheme was initially applied on flagged ENRON data in previous chapter as shown in Figure 7.1

Universiti Utara Malaysia

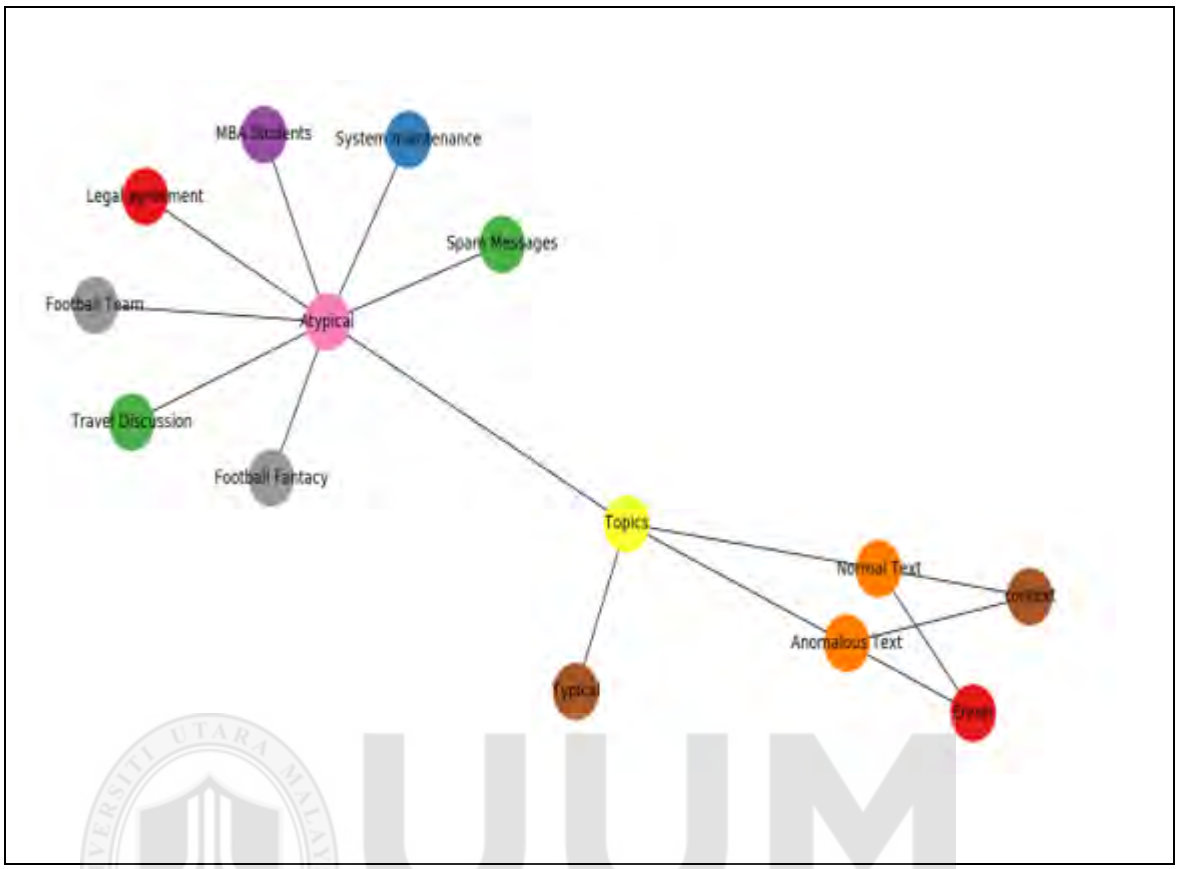

Figure 7.1. Representing ENRON data using ESET3

The aim of representing semantic base text anomalies is to lessen the challenges of understanding information that emerges from large numbers of text documents.To have a general idea of how FOL representation scheme works, Concept Network Graph was also applied on ESET1 for interpretation. Table 7.1 and Figure 7.2 shows how the detected semantic-based text anomalies were represented as POI in ESET1 to mine useful information from ENRON mail messages. In this case, frequently used words were applied on ESET1 to check relatedness of words -to- POI. Some of the identified list of terms from selected POI are as thus;

- Terms1: Business, Communication, Information Trading
- Terms2: Electricity, Gas Power, Message
- Terms3: Enron Company Email Meeting

- Terms4: Trading Market Meeting Million
- Terms5: Legal, Agreement, Form
- Terms6: Capacity storage power gas
- Terms7: Communication Mail Information Messages
- Terms8: Subject Forwarded Messages Email
- Terms9: Forms Deal Counterpart Enron
- Terms10: Credit Million Enron.Com Information
- Terms11: Information Email Mail Enron.Com
- Terms12: Agreement Form Legal Meeting

Moving on now to consider the generated terms which were simplified into queried sentence as proposed by Silveira & Branco, (2012) who made use of a two-phased terms to summarized sentenced using a Natural language processing Library called (TextBlob sentence summarization). This is solely aimed at identifying meanings to generated terms as thus;

Universiti Utara Malaysia

- Query1: Who oversaw business communication information and trading?
- Query2: Who managed messages on gas power and electricity?
- Query3: Who were the ENRON company staff in charge of sending frequent emails for meeting?
- Query4: Who goes for meetings that generated millions in the market?
- Query5: POIs in legal forms and agreement?
- Query6: Overseers of marketing power and electricity to dynergy organisation?
- Query7: POIs that deals with storing electrical power, Gas and in what capacity?
- Query8: POIs with most communicated information mail messages?

- Query9: Who are the Subjects of emails forwarded messages?
- Query10: POIs forms dealing with ENRON counterparts?
- Query11: POIs with the most email information on enron.com?
- Query12: POIs overseeing credit information worth millions on enron.com?

Table 7.1. shows value ranging from [-1,1] were scaled.

Table 7.1.

Scorecard for POIs

Query	**K. Lay**	**J.Skilling**	**R.**	**K. Mann**	**J.**	**T. Jones**	**S. Kean**	**S. Sara**	**J. Steffes**	**M. Taylor**	**D. Pete**	**C.**	**K. Syemes**
Query1	1.000	0.959	0.450	0.937	0.649	0.943	0.553	0.906	0.999	0.948	0.966	0.951	0.786
Query2	0.822	0.949	-0.137	0.968	0.101	0.964	-0.018	0.985	0.832	0.960	0.940	0.957	0.295
Query3	0.905	0.748	0.787	0.701	0.910	0.714	0.854	0.640	0.897	0.723	0.766	0.730	0.974
Query4	0.403	0.128	0.998	0.061	0.957	0.078	0.985	-0.020	0.386	0.091	0.155	0.102	0.882
Query5	0.720	0.887	-0.293	0.916	-0.058	0.909	-0.178	0.946	0.733	0.903	0.874	0.899	0.139
Query6	0.328	0.048	0.991	-0.019	0.931	-0.001	0.968	-0.101	0.311	0.011	0.075	0.022	0.841
Query7	0.874	0.975	-0.039	0.988	0.198	0.985	0.079	0.997	0.882	0.983	0.969	0.981	0.387
Query8	0.967	0.999	0.209	0.995	0.435	0.996	0.323	0.983	0.971	0.997	0.999	0.998	0.604

Universiti Utara Malaysia

Query9	0.999	0.946	0.488	0.922	0.681	0.928	0.588	0.887	0.998	0.933	0.954	0.937	0.812
Query10	0.759	0.912	-0.239	0.937	-0.002	0.931	-0.122	0.963	0.770	0.926	0.900	0.922	0.195
Query11	0.967	0.999	0.211	0.995	0.437	0.996	0.325	0.983	0.972	0.997	0.999	0.998	0.605
Query12	0.720	0.887	-0.293	0.916	-0.058	0.909	-0.178	0.946	0.733	0.903	0.874	0.899	0.139

ESET was able to categorize figure ranging from [-1, 0.2] as low and [0.3, 0.5] as average while [0.6, 1] as high.it was also noticed that many POIs share similar activities like Richard Shapiro and Sara Shackleton deals with power and gas more than other POIs. Information generated from these POIs can be used to form Cooncept Network graphs as shown in Figure 7.2 and 7.3

Universiti Utara Malaysia

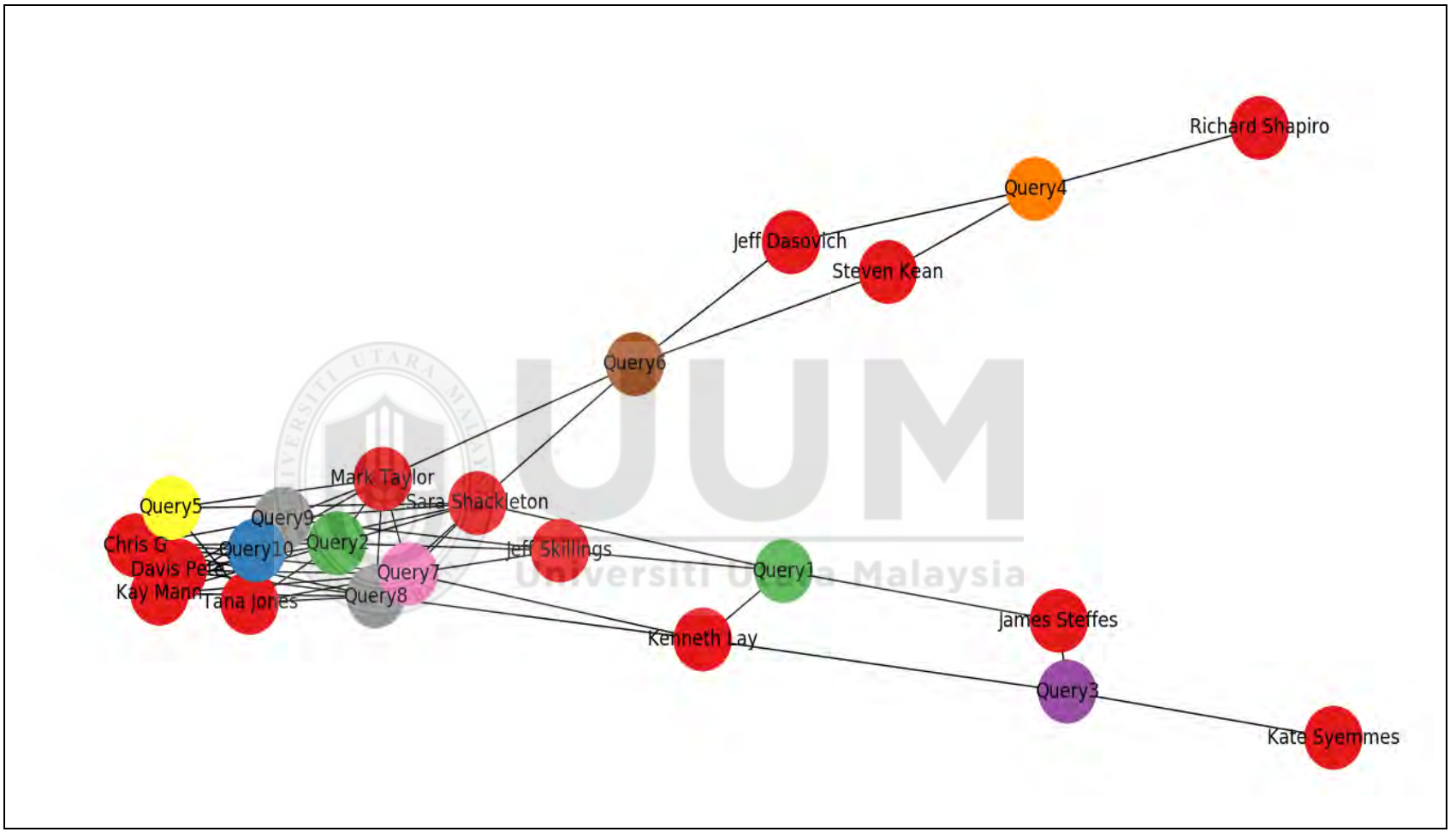

Figure 7.2. POI job connectivity

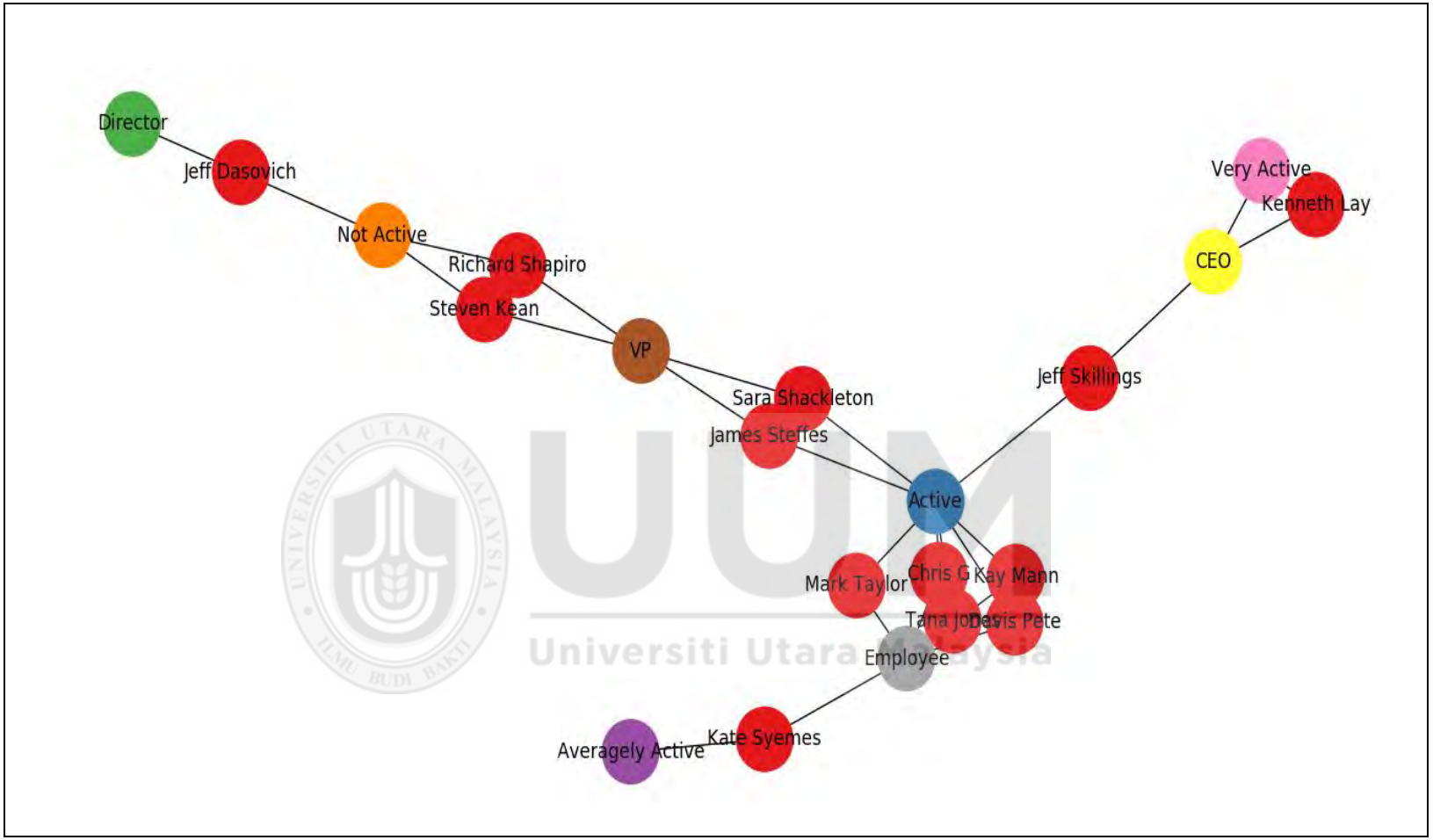

Figure 7.3. Concept Network Graph illustrating the ENRON POIs

Figure 7.2 shows individual POI with related job performance. The constructed graph can also be Departmental Network graph of individual POIs. The Graph shows series of nodes with varying colours such as Red nodes signifies 'POI', Green means 'Query1 and 2', Purple implies 'Query3', Orange implies 'Query 4', Yellow means 'Query5', Brown implies 'Query6', Pink means 'Query7', Ash is 'Query8', Ash is 'Query9' and Blue is 'Query10'.

As shown in both Figure 7.2 and 7.3 respectively, ESET was employed to create a community network graph amongst POIs. The community network graphs were created using frequency of words from the identified POIs mail messages. This was aimed at understanding the relationships of POIs departments and how active each POIs were with their job performance. The community network graph was performed by mainly comparing similarity ratio of every generated query with each POIs to check for similarity. This has helped in producing relationship as was shown in both Figure 7.2 and 7.3 respectively. Colour coding was used to convey information as thus listed; Red nodes indicates '*POI'*, Green stands for '*Director'*, Blue means '*Active Job Performer'*, Brown is '*VP'*, Orange is '*Not Active Job Performer'*, colour Ash is used to describe '*Employee'*, Pink means '*Very Active Job Performer'* and colour Purple means '*Averagely Active job Performer'.* Judging from table 7.1 value ranging from [-1,1] were scaled. ESET was able to categorize figure ranging from [-1, 0.2] as low and [0.3, 0.5] as average while [0.6, 1] as high.

Universiti Utara Malaysia

First Order Logic (FOL)

The developed standard FOL rule recognizes terms like individual variables, individual constants and predicates for instance; "*Kenneth Lay Owns"* may be formalized as Owns (*Kenneth Lay*), Owns can be referred to as Unary predicate. Semantically, this can be modelled as a set of pairs, where the proposition is true in an event where ordered pair belongs to a set. An alternative approach is by treating functions in novel styles of representation for instance; "*Kenneth Lay Owns ENRON"* can be formalized as *owns(j)(m)* rather than modelling relations. *Owns* in the given instance denotes a function which applies to one argument to yield a new function that is then applied to the second argument. Predications will be treated syntactically as functions applied to be represented as *n*-array relations and graphically phase.

FOL representation of identified terms from figure 7.4 with POIs using the Quantification scope ambiguity to resolve ambiguated concepts and as well to simplify the network graphical representation schemes to detect semantic-based text anomaly in documents. From the FOL, more meaningful information was generated using the FOL representation scheme from the identified POIs. The reason for transformation into FOL representation form is to ease the understanding and analysis compared to other forms.

{'ACTIVE': 'a2','CEO': 'a1','DIRECTORS': 'c1','EMPLOYEES': 'd1','NOT ACTIVE': 'a3', 'VP': 'b1'}
'POI': {('Kenneth Lay','Jeff Skillings', 'Richard Shapiro','Tana Jones','Kay Mann','Mark Taylor', 'Chris G','Kate Syemmes', 'Davis Pete','Steven Kean','Sara Shackleton', 'James Steffes','Jeff Dasovich')},
CEO': {('Kenneth Lay',), ('Jeff Skillings',)}, 'Chris G': {('a3',), ('d1',)},
'DIRECTORS': {('Jeff Dasovich',)}, 'Davis Pete': {('a2',), ('d1',)},
'James Steffes': {('b1',), ('a3',), ('d1',)}, 'Jeff Dasovich': {('c1',), ('d1',)},
'Jeff Skillings': {('a2',), ('a1',), ('d1',)},'Kate Syemmes': {('a3',), ('d1',)},

'Kay Mann': {('a2',), ('d1',)}, 'Kenneth Lay': {('a2',), ('a1',), ('d1',)}, 'Mark Taylor': {('a2',), ('d1',)},
'Richard Shapiro': {('b1',), ('a3',)},
'Sara Shackleton': {('b1',), ('d1',)},
'Steven Kean': {('b1',), ('d1',)},
'Tana Jones': {('a2',), ('d1',)},
'VP': {('James Steffes',), ('Richard Shapiro',),('Sara Shackleton',), ('Steven Kean',)}}
{'Mark Taylor', 'a2', 'b1', 'Tana Jones', 'Davis Pete', 'a3', 'Chris G', 'Richard Shapiro', 'James Steffes', 'Kate Syemmes', 'Sara Shackleton', 'Steven Kean', 'Kenneth Lay', 'a1', 'Jeff Skillings', 'd1', 'c1', 'Jeff Dasovich', 'Kay Mann'}
([x], [CEO(a1), EMPLOYEES(d1)])
([x], [CEO(a1), EMPLOYEES(d1)])
(([x], [CEO(a1), EMPLOYEES(d1)]) + ([y] ,[ACTIVE(a2), NOTACTIVE(a3)]))
([x, y], [CEO(a1), EMPLOYEES(d1), ACTIVE(a2), NOTACTIVE(a3)])
([],[(([x],[CEO(a1),EMPLOYEES(d1)])->([y],[ACTIVE(a2), NOTACTIVE(a3)]))])
exists x.(CEO(a1) & EMPLOYEES(d1))
exists x.(CEO(a1) & ACTIVE(a2))
| CEO(a1) | | ACTIVE(a2) |
| EMPLOYEES(d1) | | NOTACTIVE(a3) |
|_____| |_____|
([x,y],[works(x,y)])
([x],[CEO(x), EMPLOYEES(x)])

Figure 7.4. FOL representation

7.3 Representation scheme for 20NG Data

It has been shown and described how queries are converted into network graphs in previous section with the ENRON data. Turning now to the 20NGS data, the level of similarity of each document shared amongst groups was identified in section (4.2.3). Queries were generated from modelled terms to give a deeper insight of what the identified semantic based anomalous data have in common and what they don't. Figure 7.5. Shows Topics in the 20NG. The Graph shows series of nodes with varying colours such as Red nodes signifies '*GROUPS'*, Green means '*COMPUTER'*, Orange implies '*RECREATION'*, *Query5'*, Brown implies '*SCIENCE'*, Ash is '*POLITICS AND RELIGION'*. Four main topics were created from terms to simplify interpretability and representation of 20NG data as thus.

Terms 1: God, Religion, Politics, People

Query 1: what document is related to Politics and religion?

Terms 2: Graphics, Electronics, System, Computer

Query 2: what document is related to Computer systems?

Terms 3: Science, People Medicine, encryption

Query 3: what document is related to science

Terms 4: Bikes, Motorcycles, dod, bikes, recreation'

Query 4: what document is related to recreation with motor bikes?

These queries was developed into concept network graph as will be shown in Figure 7.5

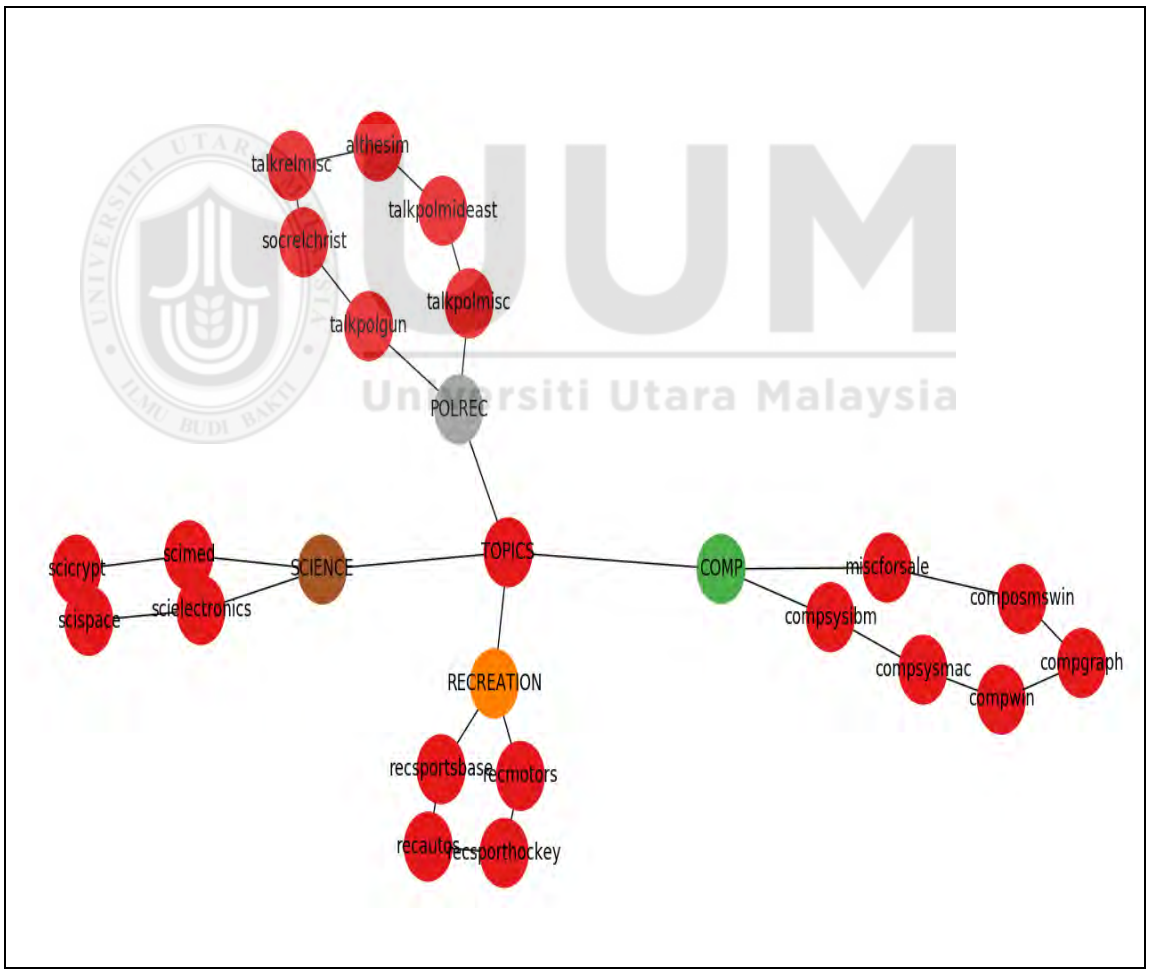

Figure 7.5. CNG representation of 20NG data using ESET3

{'TOPICS':'a2',{(althesim,),(sci.crypt,),(talk.politics.guns,),(comp.windows,),(compgraph),(rec.sport.baseball),(rec.autos),(comp.os.mswindows.misc),(talk.religion.misc),(rec.sport.hockey),(comp.sys.mac.hardware),(sci.electronics),(soc.religion.christian),(misc.forsale),(sci.med),(compsysibm),(sci.space),(rec.motorcycles),(talk.religion.misc),(talk.politics.mideast)}}

'POL-REL':'a1',{althesim,talk.politics.guns,talk.religion.misc,talk.politics.mideast, 'talkpolmisc', soc religionchrist }
'RECREATION':'c1',{recsporthockey,recautos,recsportbaseball,recmotorcycles }
'SCIENCE': 'd1', {sci.med, sci.spcae, sci.electronics, sci.crypt }
'COMPUTER': 'a3', { comp.os.mswindows.misc, compgraph, comp.windows, comp.os.mswindows.misc , compsysibm , comp.sys.mac.hardware }

'20NG':[,{(althesim,'a1','a2'),(sci.crypt,'a2','d1',),(talk.politics.guns,'a1','a2',),(comp.windows,'a2','a3',),(compgraph,'a2','a3'),(rec.sport.baseball,'a2','c1'),(rec.autos,'a2','c1'),(comp.os.mswindows.misc,'a2','a3),(talk.religion.misc,'a1','a2'),(rec.sport.hockey,'a2','c1'),(comp.sys.mac.hardware,'a2','a3),(sci.electronics,,'a2','d1'),(soc.religion.christian,'a1','a2'),(misc.forsale,'a2','a3),(sci.med),(compsysibm'a2','a3),(sci.space,'a2','d1'),(rec.motorcycles,'a2','c1'),(talk.religion.misc,'a1','a2'),(talk.politics.mideast,'a1','a2')}

([x], [TOPIC(a2), SCIENCE(d1)])
([x], [TOPIC(a2), RECREATION(c1)])
([x], [TOPIC(a2), COMPUTER(a3)])
([x], [TOPIC(a2), POL-RELE(a1)])

Figure 7.6. FOL representation of the Concept Network Graph for 20NG

Figure 7.6 Illustrates the reason for transformation 20NG into Topics using FOL representation form is to ease the understanding and analysis compared to other forms.

7.4 Representation scheme for NIPS and Daily Kos Data

ESET3 was used to detect semantic-based text anomaly by identifying Topic divergence from NIPS and Daily Kos data respective. This is aimed at identifying side information using the context and content-based topic detection approach as was seen in Mahapatra et al., (2012). This section helps in interpreting the detection trend using FOL representation scheme for visualization and understandability of text documents as shown in Figure 7.7.

and Figure 7.8.

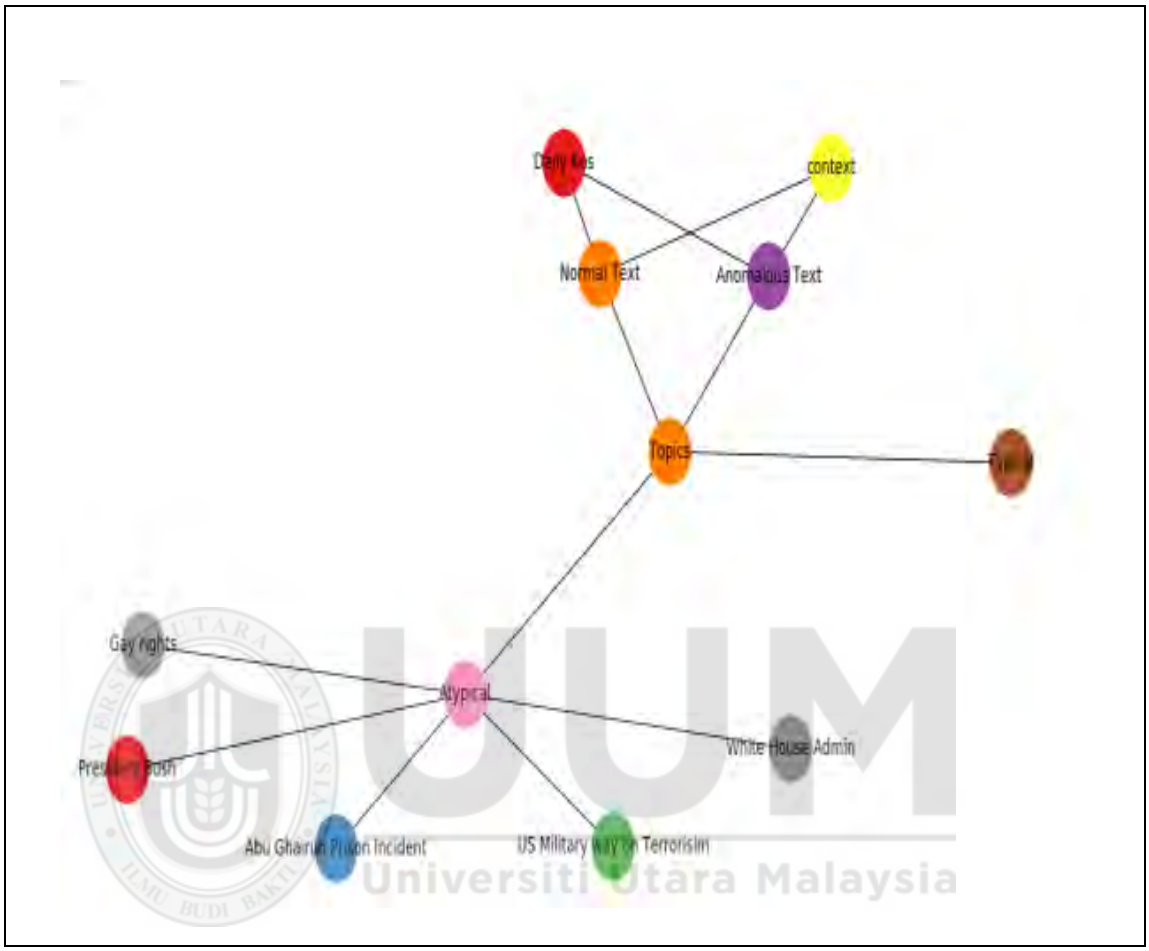

Figure 7.7. CNG representation of KOS data using ESET3

Universiti Utara Malaysia

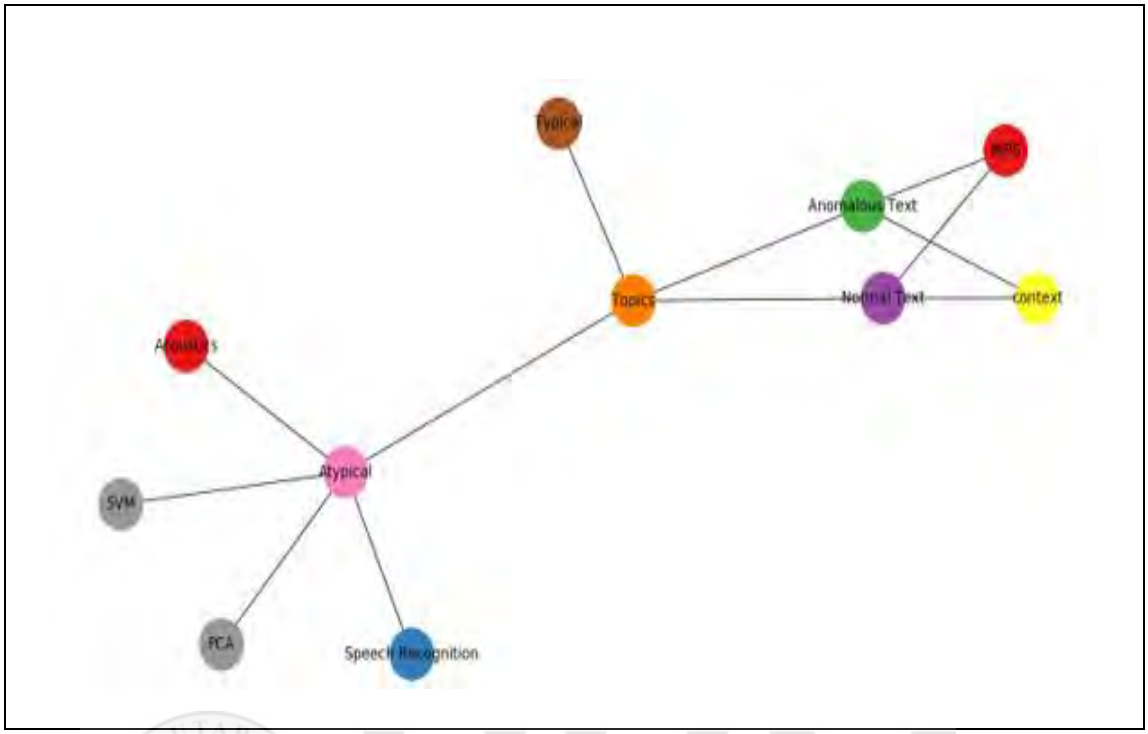

Figure 7.8. CNG representation of NIPS data using ESET3

Similarity of papers in NIPS depends fully on the text content of each papers. Ten randomly selected papers were used from NIPS web link *https://papers.nips.cc/book/advances-in-neural-information-processing-systems-26-2013* to detect and show anomaly-based text semantic similarity as used in the study as will be shown in Figure 7.9.

File1: Self organization of associative database and its application File2: A mean field theory of layer iv of visual cortex and its application to Artificial Neural Networks File3: Storing Covariance by the Associative long-term potentiation and depression File4: Bayesian Query construction for Neural Networks model File5: Neural Network to insinuate a deformable model File6: Neural Network ensemble, cross validation and activation File 7: Plasticity-mediated competitive learning File 8: ICEG morphology classification using analogue VLSI neural networks. File9: Realtime control of Tokamak plasma using Neural Networks File10: Learning to play the game of chess.

Figure 7.9. file names from NIPs conference

This randomly selected papers were named file1 to file10. All files have varying topics, date, abstract, contents and authors. To represent the anomaly-based text semantic in these files, semantic similarity of words was identified using the ESET as shown in Figure 7.10.

```
[File1[ 1.   0.87 0.75 0.85 0.87 0.88 0.85 0.85 0.87 0.88]
 File2[ 0.87 1.   0.77 0.85 0.87 0.88 0.83 0.88 0.88 0.9 ]
File3 [ 0.75 0.77 1.   0.74 0.73 0.74 0.75 0.77 0.75 0.77]
File4 [ 0.85 0.85 0.74 1.   0.88 0.89 0.82 0.87 0.86 0.88]
File5 [ 0.87 0.87 0.73 0.88 1.   0.91 0.82 0.89 0.89 0.9 ]
File6 [ 0.88 0.88 0.74 0.89 0.91 1.   0.84 0.91 0.91 0.92]
File7 [ 0.85 0.83 0.75 0.82 0.82 0.84 1.   0.82 0.82 0.84]
File8 [ 0.85 0.88 0.77 0.87 0.89 0.91 0.82 1.   0.91 0.93]
File9 [ 0.87 0.88 0.75 0.86 0.89 0.91 0.82 0.91 1.   0.98]
File10 [ 0.88 0.9  0.77 0.88 0.9  0.92 0.84 0.93 0.98 1.  ]]
```

Figure 7.10. Optimized cosine similarity of files from NIPS

Figure 7.10. shows the similarity amongst files from the NIPs conference papers. It was noticed that file 3 has a different reading entirely from other files. This indicates that the word content in file3 may not be related or may not have many discussions about *Neural Networks* (Neural network is the most semantically related word found in all ten files) compared with other files. It clearly shows that file3 with title "*Storing Covariance by the Associative long-term potentiation and depression"* is different from other files. It was noticed that this file contains entirely different information with other files. To experiment deeper on all files with NIPs data. A comprehensive network graph was able to differentiate contents from files according to data attributes as shown in Figure 7.11.

Universiti Utara Malaysia

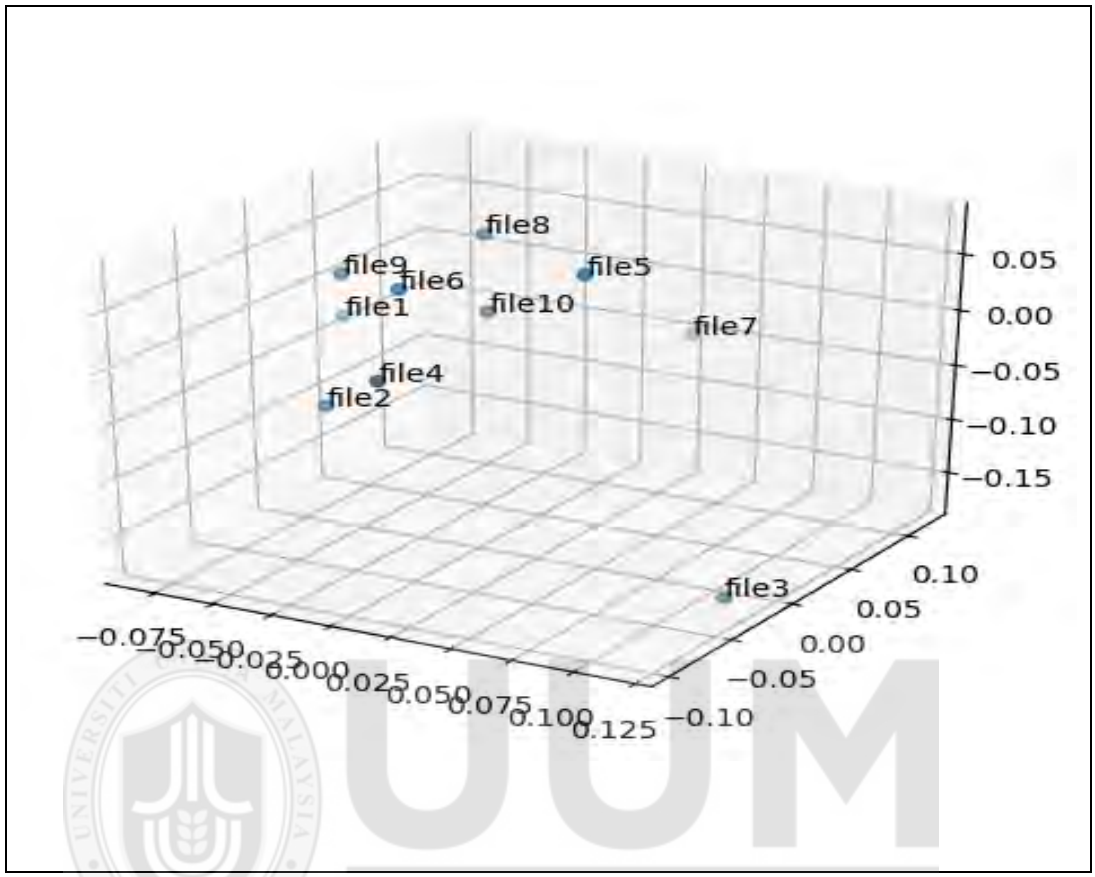

Figure 7.11. 2D graph representation of optimized cosine similarity of files from NIPS.

Figure 7.12 and 7.13 shows how NIPS text data is represented using the representation schemes.

Universiti Utara Malaysia

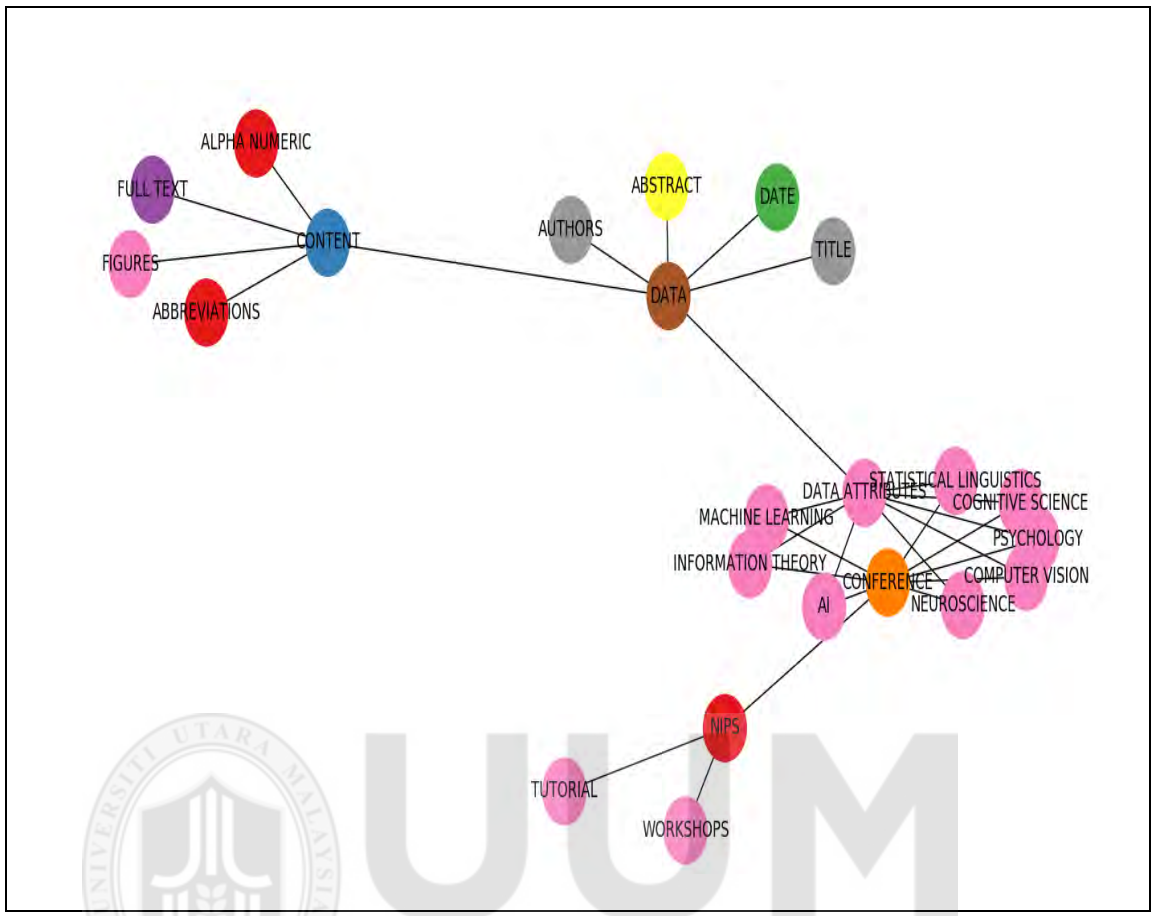

Figure 7.12. CNG in NIPs conference paper based on varying themes of information

Universiti Utara Malaysia

{‘NIPS Category’:’a1’,{(Conferences),(Tutorials),(Workshops)}
{'Conference Types':'a2',{(Machine Learning,), (Statistical Linguistics),
(Neurosciences,), (comp.windows,),(AI),(ComputerVision), (Cognitive Science),
(Information Theory), (Psychology)}}
{Data Attributes:'a3',{(Dates,), (Authors), (Abstract,), (Title), (Content)}}
{Contents:'a4',{(Full text,), (alphanumeric), (Figures), (Abbreviation)}}

‘Dates’: {('a3',), ('b1',)},
‘figures’: {('a4',), ('b2',)},
‘AI’: {('a1',), ('b5',)},
([x], [NIPS Category(a1), Contents (a4)])

([x], [NIPS Category(a1), RECREATION(c1)])

Figure 7.13. FOL representation of NIPS

CNG was able to categorize anomaly detection into levels as was described in chapter two (section 2.4.1). Another significant aspect of the CNG was its ability to create a

Community Network Graph amongst ERON POI as shown in section (7.2). while FOL was able to show information completeness from all identified semantic based text anomalies.

7.5 Summary

This chapter shows the semantic-based text anomalies employing FOL and Concept Network Graph to represent and visualize meaningful information from large corpus. The idea is based mainly to simplify the understandability and interoperability of the ESET employing representation scheme and as well to give clearer outcomes of the computed functions in ESET.

Universiti Utara Malaysia

CHAPTER EIGHT

DISCUSSION AND CONCLUSION

8.1 Introduction

This chapter finishes up the study by highlighting major research contributions, the relevance of the study to text mining community and the implication for the semantic-based text anomaly detection with future directions that can be pursued.

8.2 The Research Summary

This study focused mainly on detecting semantic-based text anomaly in documents. The approach employs anomaly detection and an Enhanced Sequential Exception Technique (ESET) was developed. Experiments were conducted to illustrate how this technique can be employed to yield promising results. ESET has significantly lessened the computational demand in detecting semantic-based text anomaly compared to existing techniques in this area. Experimental performance also shows that the technique identifies meaningful information from text documents with an improved representation scheme which adopts the combination of both First Order Logic and Concept Network Graph.

8.3 Research Contributions

In this study a novel technique has been employed to enable the detection of semantic-based text anomaly in documents. More so, an experimental design was developed for the study to illustrate the components and phases of ESET. This section summarizes the main contribution of the study by referring to the research objectives stated in section 1.5 as

thus:

a. The combined Text pre-processing method (NLP) with SET functions into forming Enhanced Sequential Exception Technique (ESET). This objective has been achieved by exploring and testing SET with text to detect anomaly. Since the existing SET focuses only on categorical data from large log files, ESET is dedicated in detecting anomalous data mainly from text data which is considered an interesting contribution to this research area. However, lack of consistent structure in text documents (heterogeneity) poses a major challenge and ESET has been successfully crafted to include various components that collaborated with each other to achieve its desired goals.

b. Optimized cosine function was employed in this study in contrast to the traditional SET which made use of variance and standard deviation on categorical data to detect anomalous data. SET made use of both variance and standard deviation, which compares high with low cardinalities of data. However, it was noticed that the optimized cosine was able to improve the detection of anomalous unstructured text data since it assumes text documents to be vectors. Vectors enable readings of word occurrence in text document by stating the rarity or frequency of text which is considered as anomaly.

c. A hybridize modified ESET with Latent Semantic Analysis was developed which carters for detection of semantic-based text anomaly in an unstructured textual data. One prominent contribution of the technique is that LSA was able to cope with the high dimensional nature of text and was able to analyse semantics in textual documents into term or concept models which achieved a significant performance.

Universiti Utara Malaysia

The adopted LSA model using SVD has low time and space complexity and can automatically infer numbers of concept or topics as models.

d. Analysis of semantic-based text anomaly in unstructured text was performed by identifying its Canonical Form (CF) using combined Word Sense Disambiguation algorithms namely; Lesk and Selectional preference. This objective is concerned with the following contribution that ensures the achievement of this objective:

- Semantic similarity matrices were initially compared structure, knowledge and corpus-based approach. This was aimed at analysing the performance of these approaches and understanding how it can improve ESET functions. It was noticed that combining these approaches would yield a better accuracy in identifying text semantic similarity in documents as well improve the detection of semantic based text anomalous data. ESET3 has a significant result over other semantic similarity measures.
- Disambiguating word meanings into preferred synonyms that carries the contextual idea or concepts of textual documents. This was performed by formulating a combined form of knowledge/corpus-based word disambiguation algorithm.
- Canonizing text into normalized forms for easy interpretability and analysis of textual documents. This enables the comparison to be lenient enough to support real world subjective text in documents.

e. The Concept Network Graph was used to create network representation which identified semantic-based text anomaly based on topic, event, sentence, document and

word as seen in chapter six respectively. This was able to categorize anomaly detection into levels as was described in chapter two (section 2.4.1). More so, it showed the functional role of every level of text anomaly detection. Another significant aspect of the Concept Network Graph was its ability to create a Community Network Graph amongst ERON POI as shown in section (7.2). Community Network Graph helps to categorize identified data into groups based on similar features or attribute shared.

f. Turning now to First Order Logic (FOL), it plays the role of predicate logic representation using logics to interpret information and creating knowledge from hidden idea. FOL plays a vital role in ESET in the sense that, it was able to represent completeness in conveyed information or identified information. FOL draws a formal expression of text validity, satisfiability and information completeness in documents (Navigli, 2009). It enables text satisfiability and validity for checkmating text semantics that was conveyed in every sentences, paragraphs and document. Texts were matched with the context of idea conveyed. If text is noticed to have a meaningful idea, complete and compliments its context of discourse, then it is evident that knowledge can be shared on that context. On the other hand, if knowledge is not shared or no meaning is derived from text context, then text information is said to have incomplete meaning. It is important to note that FOL with Concept Network Graph was able improve understandability problem as well increase searchability or detection of semantic-based text anomaly with ease from huge numbers text documents.

Thus far, the novelty of the study is in the ESET which has components or phases that provided solution to detecting semantic-based text anomaly. Empirical evaluation was mainly focused on measuring the accuracy of the detected semantic-based text anomalies which was solely based on the provided data with the benchmarked studies. Overall, the study has achieved its intended objectives outlined in chapter I. Moreover, techniques presented in this study is considered extremely suitable for mining and detecting semantic-based text anomaly from large documents.

8.4 Future Work

Although the experimental results were proven to be favourable as the study objectives. Nevertheless, there were several limitations of this research work and it provides fertile ground for future research. For the Latent semantic analysis phase with ESET1 (ESET2). Despite its success, there were some noticeable issues or shortcomings in ESET2 especially in the LSA part. Existing researches have improved standard LSA employing other ways like the distribution, probabilistic method, constraining sparsity and regularization. The weakness in LSA relating to density processing of orthogonal matrices, complexities in decomposing matrix, problems faced with alternative iteration algorithms needs to be improved especially when dealing with very large document data. Another noticeable weakness in the study is ESET3 using the combined or integrated WSD algorithms. It was noticed that WSD algorithms works perfectly on list of words better than sentences. To produce a better result, another algorithm should be made that disambiguates sentences from large number of documents. This will properly check the hierarchical relatedness of words and coin its meaning to properly detect semantic-based text anomalous data in large numbers of documents as presaged by this study.

Universiti Utara Malaysia

Nevertheless, the study had some setbacks in identifying terms like names, abbreviations and special keywords or terms. This needs to be considered when trying to detect text anomalies especially in corpus with special keywords. Nonetheless, there was a noticeable improvement with the use of the combined WSD algorithms to canonize text and resolve synonym and polysemous related problems.

Recently, researchers are beginning to show interest in detecting fake news, identifying filter bubbles as well using an unsupervised approach to swiftly detect semantic based text anomalies in documents. However, detection of semantic based text anomalies may seem tasking for huge numbers of documents. For text representation, text representation using the concept network with FOL was used as a representation scheme. Semantic-based text anomaly where possible to interpret with full textual data since concepts and information can be easily read. However, BOW data prove to be difficult to easily represent and identify knowledge. Therefore, some data were represented in the study instead of all data. A possible solution is to device a machine learning algorithm that will be able to work perfectly on full textual data and BOW as also needed in ESET3.

Universiti Utara Malaysia

Besides the suggestions, it is irrefutable that further work can improve ESET significantly. Finally, this technique can be further explored to support future research in machine learning, text mining, data mining, field of algorithm development and computer science.

References

A.Rajaraman, J. Leskovec, J. D. U. (2016). *Mining Massive Data Sets Winter 2016*. Cambridge University Press. Retrieved from http://web.stanford.edu/class/cs246

ABDULSAHIB, A. K. (2015). *Graph based text representation for document clustering asma khazaal abdulsahib*.

Abdulsahib, A. K., & Kamaruddin, S. S. (2015). Graph based text representation for document clustering. *Journal of Theoretical and Applied Information Technology*, *76*(1), 1–13. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-84930694414&partnerID=40&md5=5c7f0059c26594915cdf9360315173c7

Abouzakhar, N., Allison, B., & Guthrie, L. (2008). Unsupervised Learning-based Anomalous Arabic Text Detection. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*, 291–296. Retrieved from http://www.lrec-conf.org/proceedings/lrec2008/summaries/83.html

Acree, B., Jansa, J., & Shoub, K. (2016). Comparing and Evaluating Cosine Similarity Scores, Weighted Cosine Similarity Scores, and Substring Matching. Retrieved from https://shoub.web.unc.edu/files/2016/04/AHJS_Weighted_Cosine.pdf

Adler-Golden, S. M. (2009). Improved hyperspectral anomaly detection in heavy-tailed backgrounds. *WHISPERS '09 - 1st Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*, 2–5. https://doi.org/10.1109/WHISPERS.2009.5289019

Aggarwal, C., & Zhai, C. (2012). *Mining text data*. (C. C. C. Z. AGGARWAL, Ed.), *Mining Text Data* (Vol. 4). Kluwer Academic Publishers Boston/Dordrecht/London. https://doi.org/10.1007/978-1-4614-3223-4

Agirre, E., & Martinez, D. (2002). Integrating selectional preferences in WordNet. *Proceedings of the First International WordNet Conference*, 9. Retrieved from http://arxiv.org/abs/cs/0204027

Akarsu, B., Bayram, K., Slisko, J., & Corona Cruz, A. (2013). International Journal Of Scientific Research And Education. *Ijsae.In*, *6*(3), 221–232. Retrieved from http://ijsae.in/ijsaeems/index.php/ijsae/article/viewFile/157/137

Akoglu, L., Tong, H., & Koutra, D. (2014). Graph-based Anomaly Detection and Description: A Survey. *ArXiv Preprint ArXiv:1404.4679*, 49. https://doi.org/10.1007/s10618-014-0365-y

Alagi, D. (2009). Experiments on Active Learning for Croatian Word Sense Disambiguation.

Allan Collins, J. S. B., Larkin, & K. M., & Newman, B. B. and. (2007). *INFERENCE IN TEXT UNDERSTANDING*. University of Illinois at Urbana- Champaign 51 Gerty Drive Champaign, Illinois 61820.

Allan, J., Carbonell, J., & Doddington, G. (1998). Topic detection and tracking pilot study: Final report. *DARPA Broadcast News Transcription and Understanding Workshop.*, 194–218. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.21.6373&rep=rep1

Universiti Utara Malaysia

&type=pdf

Almarimi, A., & Andrejková, G. (2016). Text Anomalies Detection Using Histograms of Words. *ACSIJ Advances in Computer Science: An International Journal*, *5*(1), 63–68.

Arning, A., & Rakesh, A. (1996). Method for Deviation in Large Databases. *KDD-96 Proceedings*.

Atefeh, F., & Khreich, W. (2015). A Survey of Techniques for Event Detection in Twitter TECHNIQUES FOR EVENT DETECTION IN TWITTER. *Computational Intelligence*, *0*(1), 132–164. https://doi.org/10.1111/coin.12017

Balbi, S. (2010). Beyond the curse of multidimentionality: high dimentional clustering in context mining. *Statistica Applicata - Italian Journal of Applied Statistics*, *22*(1), 53–63.

Banerjee, S. (2002). Adapting the Lesk Algorithm for Word Sense Disambiguation to WordNet, (December).

Basile, P., Caputo, A., & Semeraro, G. (2014). An Enhanced Lesk Word Sense Disambiguation Algorithm through a Distributional Semantic Model. *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers (COLING 14)*, 1591–1600.

Belford, M., Mac Namee, B., & Greene, D. (2018). Stability of topic modeling via matrix factorization. *Expert Systems with Applications*, *91*, 159–169. https://doi.org/10.1016/j.eswa.2017.08.047

Universiti Utara Malaysia

Beltagy, I., Roller, S., Cheng, P., Erk, K., & Mooney, R. J. (2015). Representing Meaning with a Combination of Logical Form and Vectors, 1–44. Retrieved from http://arxiv.org/abs/1505.06816

Berant, J., Chou, A., Frostig, R., & Liang, P. (2013). Semantic Parsing on Freebase from Question-Answer Pairs. *Proceedings of EMNLP*, (October), 1533–1544. Retrieved from https://www.aclweb.org/anthology/D/D13/D13-1160.pdf%5Cnhttp://www.samstyle.tk/index.pl/00/http/nlp.stanford.edu/pubs/semparseEMNLP13.pdf

Bernotas, M., Karklius, K., Laurutis, R., & Slotkiene, A. (2007). The peculiarities of the text document representation, using ontology and tagging-based clustering technique. *Information Technology and Control*, *36*(2), 217–220.

Bertoldi, N., Cettolo, M., & Federico, M. (2010). Statistical Machine Translation of Texts with Misspelled Words. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, (June), 412–419.

Bhaduri, K., Matthews, B. L., & Giannella, C. R. (2011). Algorithms for speeding up distance-based outlier detection. *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 859–867. https://doi.org/10.1145/2020408.2020554

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2012). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(4–5), 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993

Universiti Utara Malaysia

Boyd-Graber, J., Blei, D. M., & Zhu, X. (2007). A Topic Model for Word Sense Disambiguation. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL'07)*, 1024–1033.

Brants, T., Chen, F., & Farahat, A. (2003). A system for new event detection. *ACM SIGIR Conference on Research and Development in Informaion Retrieva*, (pp. 330-337).

Brants, T., Chen, F., & Tsochantaridis, I. (2002). Topic-based document segmentation with probabilistic latent semantic analysis. *Proceedings of the Eleventh International Conference on Information and Knowledge Management CIKM 02*, 211. https://doi.org/10.1145/584792.584829

Breja, M. (2015). A Novel approach for Novelty Detection of Web Documents, *6*(5), 4257–4262.

Brody, S. (2005). Cluster-Based Pattern Recognition in Natural Language Text. *English*, (August). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.81.7288&rep=rep1&type=pdf

Bruynooghe, M., & Denecker, M. (2014). First Order Logic with Inductive Definitions for Model-Based Problem Solving.

Bustince, H., Fernadez, J., & Mesiar, R. (2011). Restricted dissimilarity functions and penalty functions. *Eusflat-Lfa 2011*, (July). Retrieved from

Universiti Utara Malaysia

http://library.utia.cas.cz/separaty/2012/E/mesiar-restricted dissimilarity functions and penalty functions.pdf

Cai, D., He, X., Wu, X., & Han, J. (2008). Non-negative matrix factorization on manifold. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 63–72. https://doi.org/10.1109/ICDM.2008.57

Cambria, E., & Melfi, G. (2015). Semantic Outlier Detection for Affective Common-Sense Reasoning and Concept-Level Sentiment Analysis, 276–281.

Cammert, M., Heinz, C., Kramer, J., & Riemenschneider, T. (n.d.). Systems and/or methods for event stream deviation detection. *U.S. Patent No. 9,659,063. Washington, DC: U.S. Patent and Trademark Office*. Retrieved from https://www.google.com/patents/US9659063

Capurro, I., Lecumberry, F., Martín, Á., Ramírez, I., Rovira, E., & Seroussi, G. (2016). Efficient sequential compression of multi-channel biomedical signals. *IEEE Journal of Biomedical and Health Informatics*, *PP*(NN), 13. Retrieved from http://arxiv.org/abs/1605.04418

Cha, S. (2007). Comprehensive Survey on Distance / Similarity Measures between Probability Density Functions, *1*(4).

Chandarana, D. R. (2015). A Survey for Different Approaches of Outlier Detection in Data Mining, 1–4.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, *41*(September), 1–58.

https://doi.org/10.1145/1541880.1541882

Chaplot, D. S., & Salakhutdinov, R. (2018). Knowledge-based Word Sense Disambiguation using Topic Models. Retrieved from http://arxiv.org/abs/1801.01900

Chen, X., & Wu, C. (2012). A Text Representation Method Based on Harmonic Series. In *IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), 2012* (pp. 1830–1834).

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, I. (2008). Text classification and Naive Bayes. Retrieved from lp.stanford.edu/IR-book/html/htmledition/text-classification-and-naive-bayes-1.html

Cichosz, P. (2018). Anomaly detection in discussion forum posts using global vectors. In *SPIE*. Proc. SPIE 10808. https://doi.org/10.1117/12.2501345

Classen, A., Boucher, Q., & Heymans, P. (2011). A text-based approach to feature modelling: Syntax and semantics of TVL. *Science of Computer Programming*, *76*(12), 1130–1143. https://doi.org/10.1016/j.scico.2010.10.005

Dang, S., & Ahmad, P. H. (2014). Text Mining : Techniques and its Application, *1*(4), 22–25.

Debortoli, S., Müller, O., Junglas, I. A., & vom Brocke, J. (2016). Text Mining for Information Systems Researchers: An Annotated Tutorial. *Manuscript Submitted for Publication*, (April).

Universiti Utara Malaysia

Deshpande, R., Vaze, K., Rathod, S., & Jarhad, T. (2014). Comparative Study of Document Similarity Algorithms and Clustering Algorithms for Sentiment Analysis. *Ijettcs.Org*, *3*(5), 196–199. Retrieved from http://www.ijettcs.org/Volume3Issue5/IJETTCS-2014-10-21-85.pdf

Ding, R., Nallapati, R., Xiang, B., & Services, A. W. (2016). Coherence-Aware Neural Topic Modeling, *1*.

Drissi, M., & Watkins, O. (2017). Hierarchical Text Generation using an Outline.

Eshghi, A., Howes, C., Gregoromichelaki, E., Hough, J., & Purver, M. (2015). *Feedback in Conversation as Incremental Semantic Update*. *Iwcs 2015*. Retrieved from http://www.aclweb.org/website/old_anthology/W/W15/W15-01.pdf#page=123

Faruqui, M., Tsvetkov, Y., Rastogi, P., & Dyer, C. (2016). Problems With Evaluation of Word Embeddings Using Word Similarity Tasks. https://doi.org/10.18653/v1/W16-2506

Foltz, P. W. (1996). Latent Semantic Analysis for Text-Based. *Behavior Research Methods, Instruments and Computers*, *28*(2), 197–202. https://doi.org/10.3758/BF03204765

Franzoni, V. (2017). Just an Update on PMING Distance for Web-based Semantic Similarity in Artificial Intelligence and Data Mining, 1–3. https://doi.org/10.13140/RG.2.2.20531.22560

Froud, H., Lachkar, A., & Ouatik, S. (2013). Arabic text summarization based on latent semantic analysis to enhance Arabic documents clustering. *ArXiv Preprint*

ArXiv:1302.1612. Retrieved from http://arxiv.org/abs/1302.1612

Furtado, P., Nadal, S., Peralta, V., Djedaini, M., & Marcel, P. (2015). Materializing Baseline Views for Deviation Detection Exploratory OLAP, 1–12.

Fyshe, A., Talukdar, P., Murphy, B., & Mitchell, T. (2013). Documents and Dependencies : an Exploration of Vector Space Models for Semantic Composition. *Conll*, 84–93.

Gabrilovich, Evgeniy, and S. M. (2005). Feature generation for text categorization using world knowledge. *IJCAI International Joint Conference on Artificial Intelligence*, *5*(pp. 1048-1053.).

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. *IJCAI International Joint Conference on Artificial Intelligence*, 1606–1611. https://doi.org/10.1145/2063576.2063865

Gahl, S., Menn, L., Ramsberger, G., Jurafsky, D. S., Elder, E., Rewega, M., & Audrey, L. H. (2003). Syntactic frame and verb bias in aphasia: Plausibility judgments of undergoer-subject sentences. *Brain and Cognition*, *53*(2), 223–228. https://doi.org/10.1016/S0278-2626(03)00114-3

Garrette, D., Erk, K., & Mooney, R. (2014). A Formal Approach to Linking Logical Form and Vector-Space Lexical Semantics. *Computing Meaning SE - 3*, *47*, 27–48. https://doi.org/10.1007/978-94-007-7284-7_3

Gelbukh, A., Sidorov, G., & Han, S.-Y. (2005). On some optimization heuristics for lesk-like WSD algorithms. *Nldb'05*, 402–405.

Universiti Utara Malaysia

Giannoulis, P., Potamianos, G., & Maragos, P. (2018). On the Joint Use of NMF and Classification for Overlapping Acoustic Event Detection. *Proceedings*, *2*(2), 90. https://doi.org/10.3390/proceedings2020090

Gilad Katz, Yuval Elovici, & B. S. (2014). *SEMANTIC BASED CONTEXTUAL CLUSTERING FOR DATA LEAKAGE PREVENTION THROUGH ANOMALY DETECTION*.

Gloor, P. A., Niepel, S., L, Y., Whalley, G., Skilling, J. K., Kitchen, L., & Causey, R. (2006). Identifying Potential Suspects by Temporal Link Analysis Discovering Suspicious Activity in the Enron e-Mail Dataset Filtering by Keywords, 9.

Godbole, S. (2002). Exploiting confusion matrices for automatic generation of topic hierarchies and scaling up multi-way classifiers. *Progress Report, IIT Bombay*, (March 2002), 17. Retrieved from http://www.it.iitb.ac.in/~shantanu/work/report.pdf

Goldstein, M., Goldstein, M., & Uchida, S. (2016). A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data. *PLoS ONE*, (April), 1–31. https://doi.org/10.7910/DVN/OPQMVF

Gomaa, W. H. (2013). A Survey of Text Similarity Approaches. *International Journal of Computer Applications*, *68*(13), 13–18.

Gong, Y., Zhao, K., & Zhu, K. Q. (2016). Representing Verbs as Argument Concepts. *Proceedings of the 30th Conference on Artificial Intelligence (AAAI 2016)*, 2615–2621.

Universiti Utara Malaysia

Goodfellow, I. (2016). NIPS 2016 Tutorial: Generative Adversarial Networks. https://doi.org/10.1001/jamainternmed.2016.8245

Guthrie, D. (2008). Unsupervised Detection of Anomalous Text. *Distribution*, (July).

Guthrie, D., Guthrie, L., Allison, B., & Wilks, Y. (2007). Unsupervised anomaly detection. *IJCAI International Joint Conference on Artificial Intelligence*, 1624–1628.

H, S. D., M, M. K., & Science, C. (2015). International Journal of Combined Research & Development (IJCRD) eISSN : 2321-225X ; pISSN : 2321-2241 Volume : 4 ; Issue : 2 ; February -2015 A Survey on Text Mining Approaches International Journal of Combined Research & Development (IJCRD), 251–256.

Han, J. (2014). Data Mining : Concepts and Techniques.

Hardin, J. S., Sarkis, G., & Urc, P. C. (2015). Network analysis with the enron email corpus. *Journal of Statistics Education*, *23*(2). https://doi.org/10.1080/10691898.2015.11889734

Hassan, S., & Mihalcea, R. (2011). Semantic Relatedness Using Salient Semantic Analysis. *Proceedings of the 25th AAAI Conference on Artificial Intelligence, (AAAI 2011)*, 884–889. Retrieved from http://www.samerhassan.com/images/4/48/Hassan.pdf%5Cnhttp://www.aaai.org/ocs/index.php/AAAI/AAAI11/paper/download/3616/3972

Héas, P., Drémeau, A., & Herzet, C. (2016). An Efficient Algorithm for Video Superresolution Based on a Sequential Model. *SIAM Journal on Imaging Sciences*,

9(2), 537–572. https://doi.org/10.1137/15M1023956

Henriksson, A., Moen, H., Skeppstedt, M., Daudaravičius, V., & Duneld, M. (2014). Synonym extraction and abbreviation expansion with ensembles of semantic spaces. *Journal of Biomedical Semantics*, *5*(1), 6. https://doi.org/10.1186/2041-1480-5-6

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, *349*(6245), 261–266. https://doi.org/10.1126/science.aaa8685

Hodge, V. J., & Austin, J. (2004). A Survey of Outlier Detection Methodoligies. *Artificial Intelligence Review*, *22*(1969), 85–126. https://doi.org/10.1007/s10462-004-4304-y

Huang, A. (2008). Similarity measures for text document clustering. *Proceedings of the Sixth New Zealand*, (April), 49–56. Retrieved from http://nzcsrsc08.canterbury.ac.nz/site/proceedings/Individual_Papers/pg049_Similarity_Measures_for_Text_Document_Clustering.pdf

Issa, H., & Vasarhelyi, M. A. (2011). Application of Anomaly Detection Techniques to Identify Fraudulent Refunds. *SSRN Working Papers Series*, 1–19. https://doi.org/10.2139/ssrn.1910468

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Janz, A., Kędzia, P., & Piasecki, M. (2018). Graph-based complex representation in inter-sentence relation recognition in Polish texts. *Cybernetics and Information*

Technologies, *18*(1), 152–170. https://doi.org/10.2478/cait-2018-0013

Jiang, L., Zhang, H., Yang, X., & Xie, N. (2013). Research on Semantic Text Mining Based on Domain Ontology, 336–343.

Joachims, T. (1998). Text Categorization with Suport Vector Machines: Learning with Many Relevant Features. *Proceedings of the 10th European Conference on Machine Learning*, 137–142. https://doi.org/10.1007/BFb0026683

Jurafsky, D., & Martin, J. H. (2000). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. *Speech and Language Processing An Introduction to Natural Language Processing Computational Linguistics and Speech Recognition*, *21*, 0–934. https://doi.org/10.1162/089120100750105975

Kamaruddin, S. S. B. (2011). *FRAMEWORK FOR DEVIATION DETECTION IN TEXT*.

Kamaruddin, S. S., Bakar, A. A., Hamdan, A. R., Nor, F. M., Nazri, M. Z. A., Othman, Z. A., & Hussein, G. S. (2015). A text mining system for deviation detection in financial documents. *Intelligent Data Analysis*, *19*(s1), S19–S44. https://doi.org/10.3233/IDA-150768

Kamaruddin, S. S., Hamdan, A. R., & Bakar, A. A. (2007). Text Mining for Deviation Detection in Financial Statement, 446–449.

Kamaruddin, S. S., Hamdan, A. R., Bakar, A. A., & Mat Nor, F. (2012). Deviation detection in text using conceptual graph interchange format and error tolerance dissimilarity function. *Intelligent Data Analysis*, *16*(3), 487–511.

Universiti Utara Malaysia

https://doi.org/10.3233/IDA-2012-0535

Kamruzzaman, S. M., Haider, F., & Hasan, A. R. (2010). Text Classification using Data Mining. *Science*, 19. Retrieved from http://arxiv.org/abs/1009.4987

Kannan, R., Woo, H., Aggarwal, C. C., & Park, H. (2017). Outlier Detection for Text Data : An Extended Version. *ArXiv*, 489–497.

Kannan, Ramakrishnan, Woo, H., Aggarwal, C. C., & Park, H. (2017). Outlier Detection for Text Data : An Extended Version. Retrieved from http://arxiv.org/abs/1701.01325

Karkali, M., Rousseau, F., Ntoulas, A., & Vazirgiannis, M. (2014). Using temporal IDF for efficient novelty detection in text streams. *ArXiv*, 30. Retrieved from http://arxiv.org/abs/1401.1456

Katariya, N. P., & Chaudhari, M. S. (2015). 126. Text Preprocessing for Text Mining Using Side Information. *International Journal of Computer Science and Mobile Applications*, *3*, 3–7.

Kim, J., & Montague, P. (2017). An Efficient Semi-Supervised SVM for Anomaly Detection, 2843–2850.

Kobus, C., Yvon, F., & Damnati, G. (2008). Normalizing SMS: are two metaphors better than one? *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, (August), 441–448. Retrieved from http://dl.acm.org/citation.cfm?id=1599137

Koehrsen, W. (2017). Machine Learning with Python on the Enron Dataset. Retrieved November 23, 2018, from https://medium.com/@williamkoehrsen/machine-learning-with-python-on-the-enron-dataset-8d71015be26d

Kshirsagar, M., Thomson, S., Schneider, N., Carbonell, J., Smith, N. a, & Dyer, C. (2015). Frame-Semantic Role Labeling with Heterogeneous Annotations. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 218–224.

Kumar, a A. (2012). Text Data Pre-processing and Dimensionality Reduction Techniques for Document Clustering Sri Sivani College of Engineering Sri Sivani College of Engineering, *1*(5), 1–6.

Kumar Palaniswamy Supervisor, H., & Aldous, D. (2015). Exploratory Data Analysis of Enron Emails.

Kumaraswamy, R., & Shavlik, J. (2012). Anomaly Detection in Text : The Value of Domain Knowledge, 225–228.

Lee, Hanjun, Keunho Choi, Donghee Yoo, Yongmoo Suh, Soowon Lee, G. H. (2017). Recommending valuable ideas in an open innovation community A text mining approach to information overload problem. https://doi.org/10.1108/eb057530

Lenci, A., Montemagni, S., & Pirrelli, V. (2001). The Acquisition and Representation of Word Meaning The Acquisition and Representation of Word Meaning . An Overview.

Universiti Utara Malaysia

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries. *Proceedings of the 5th Annual International Conference on Systems Documentation - SIGDOC '86*, 24–26. https://doi.org/10.1145/318723.318728

Leveling, J. (2007). IRSAW – Towards Semantic Annotation of Documents for Question Answering.

Leyzerov, O. (2017). Identifing Fraud from Enron Email and financial data. Retrieved November 23, 2018, from https://olegleyz.github.io/enron_classifier.html

Li, L., Hu, X., Hu, B. Y., Wang, J., & Zhou, Y. M. (2009). Measuring sentence similarity from different aspects. *Proceedings of the 2009 International Conference on Machine Learning and Cybernetics*, *4*(July), 2244–2249. https://doi.org/10.1109/ICMLC.2009.5212182

Li, L. I. N., Hu, X. I. A., Hu, B., Wang, J. U. N., & Zhou, Y. (2009). MEASURING SENTENCE SIMILARITY FROM DIFFERENT ASPECTS, (July), 12–15.

Li, X., Member, D. F., Croft, W. B., Head, D., & University, B. E. T. (2006). Sentence Level Information Patterns for Novelty Detection, 1–10. https://doi.org/10.1145/1183614.1183652

Liang, H., Tsai, F. S., & Kwee, A. T. (2009). Detecting novel business blogs. *ICICS 2009 - Conference Proceedings of the 7th International Conference on Information, Communications and Signal Processing*. https://doi.org/10.1109/ICICS.2009.5397541

Lin, Y.-S., Jiang, J.-Y., & Lee, S.-J. (2014). A Similarity Measure for Text

Classification and Clustering. *IEEE Transactions on Knowledge and Data Engineering*, *26*(7), 1575–1590. https://doi.org/10.1109/TKDE.2013.19

Liu, H., Ke, W., Wei, K. K., & Hua, Z. (2013). The impact of IT capabilities on firm performance: The mediating roles of absorptive capacity and supply chain agility. *Decision Support Systems*, *54*(3), 1452–1462. https://doi.org/10.1016/j.dss.2012.12.016

Liu, Z. (2013). *High Performance Latent Dirichlet Allocation for Text Mining*.

M. J. Denny & A. Spirling. (2018). Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It.

Mahapatra, A., Srivastava, N., & Srivastava, J. (2012). Contextual anomaly detection in text data. *Algorithms*, *5*(4), 469–489. https://doi.org/10.3390/a5040469

Maitra, Anutosh (Bangalore, I., Mohamedrasheed, Annervaz Karukapadath (Trichur, I., Jain, Tom Geo (Bangalore, I., Shivaram, Madhura (Bangalore, I., Sengupta, Shubhashis (Bangalore, I., Ramnani, Roshni Ramesh (Bangalore, I., … Sahu, Vedamati (Bangalore, I. (2016). SYSTEM FOR AUTOMATED ANALYSIS OF CLINICAL TEXT FOR PHARMACOVIGILANCE. Retrieved June 17, 2016, from http://www.freepatentsonline.com/y2016/0048655.html

Manevitz, L. M. (2001). One-Class SVMs for Document Classification. *Journal of Machine Learning Research*, *2*, 139–154. https://doi.org/10.1162/15324430260185574

Margaret Rouse. (2005). First order predicate Logic. Retrieved October 3, 2015, from

http://whatis.techtarget.com/definition/first-order-logic

Marvin, R. (2018). Exploring Word Sense Disambiguation Abilities of Neural Machine Translation Systems, *1*, 125–131.

McInnes, B. T., & Pedersen, T. (2013). Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text. *Journal of Biomedical Informatics*, *46*(6), 1116–1124. https://doi.org/10.1016/j.jbi.2013.08.008

Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., & Hurdle, J. F. (2008). Extracting Information from Textual Documents in the Electronic Health Record: A Review of Recent Research. *IMIA Yearbook of Medical Informatics Methods Inf Med*, *47*(1), 128–144. https://doi.org/me08010128

Mihalcea, R., Corley, C., & Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. *Proceedings of the 21st National Conference on Artificial Intelligence*, *1*, 775–780. https://doi.org/10.1.1.65.3690

Miller, R. C., & Myers, B. A. (2001). Outlier finding. *Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology - UIST '01*, 81. https://doi.org/10.1145/502348.502361

Montes-y-gómez, M., Gelbukh, A. F., & López-lópez, A. (2002a). Detecting Deviations in Text Collections: An Approach Using Conceptual Graphs. *Mexican International Conference on Artificial Intelligence*, 176–184. https://doi.org/10.1007/3-540-46016-0_19

Montes-y-gómez, M., Gelbukh, A., & López-lópez, A. (2002b). Text Mining at Detail

Level Using Conceptual Graphs, 122–136.

Nakov, P. (2013). On the interpretation of noun compounds: Syntax, semantics, and entailment. *Natural Language Engineering*, *19*(03), 291–330. https://doi.org/10.1017/S1351324913000065

Navigli, R. (2009a). Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, *41*(3), 10. https://doi.org/10.1145/1459352.1459355

Navigli, R. (2009b). Word sense disambiguation. *ACM Computing Surveys*, *41*(2), 1–69. https://doi.org/10.1145/1459352.1459355

Ngai, E. W. T., Hong, T., Polytechnic, K., Hom, H., Kong, H., Hom, H., & Kong, H. (2016). a Review of the Literature on Applications of Text Mining in Policy Making.

Oberreuter, G., & Velásquez, J. D. (2013). Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style. *Expert Systems with Applications*, *40*(9), 3756–3763. https://doi.org/10.1016/j.eswa.2012.12.082

Otterbacher, J., & Radev, D. (2006). Fact-focused novelty detection: A feasibility study. *Proceedings of the Twenty-Ninth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, *2006*, 687–688. https://doi.org/10.1145/1148170.1148318

Pappas, Y. (2018). Fraud Detection Using Machine Learning (Analysis). Retrieved November 23, 2018, from http://www.yannispappas.com/Fraud-Detection-Using-Machine-Learning/

Parr, T. (2012). jguru. Retrieved January 1, 2015, from http://www.jguru.com/faq/view.jsp?EID=81

Patel, F. N., & Soni, N. R. (2012). Text mining: A Brief survey. *International Journal of Advanced Computer Research*, *2*(6), 243–248. Retrieved from http://www.theaccents.org/ijacr/papers/conference/icett2012/43.pdf

Pawar, A. M. (2015). A Comprehensive Survey on Online Anomaly Detection, *119*(17), 41–45.

Peter Norvig. (2015). Natural Language Processing What We Do. Retrieved December 9, 2015, from http://research.google.com/pubs/NaturalLanguageProcessing.html

Poon, H., & Domingos, P. (2010). Unsupervised ontology induction from text. *Proceedings of the 48th Annual Meeting of the …*, (July), 296–305. Retrieved from http://dl.acm.org/citation.cfm?id=1858712

Powers, D. M. W. (2015). What the F-measure doesn't measure: Features, Flaws, Fallacies and Fixes. https://doi.org/KIT-14-001

Pradhan, N., Gyanchandani, M., & Wadhvani, R. (2015). A Review on Text Similarity Technique used in IR and its Application. *International Journal of Computer Applications*, *120*(9), 29–34. https://doi.org/10.5120/21257-4109

Provost, F., Fawcett, T., & Kohavi, R. (1997). The Case Against Accuracy Estimation for Comparing Induction Algorithms. *Proceedings of the Fifteenth International Conference on Machine Learning1*, 445–453.

Universiti Utara Malaysia

Ramage, D., Heymann, P., Manning, C. D., & Garcia-Molina, H. (2009). Clustering the tagged web. *Proceedings of the Second ACM International Conference on Web Search and Data Mining - WSDM '09*, 54. https://doi.org/10.1145/1498759.1498809

Ramya, R. S., Venugopal, K. R., Iyengar, S. S., & Patnaik, L. M. (2016). Feature Extraction and Duplicate Detection for, *16*(5).

Ray, S., & Craven, M. (2001). Representing sentence structure in hidden Markov models for information extraction. *International Joint Conference On*, *17*(1), 1273–1279. Retrieved from http://scholar.google.com/scholar?q=intitle:Representing+Sentence+Structure+in+Hidden+Markov+Models+for+Information+Extraction#0

Ren, F., & Sohrab, M. G. (2013). Class-indexing-based term weighting for automatic text classification. *Information Sciences*, *236*, 109–125. https://doi.org/10.1016/j.ins.2013.02.029

Rennie, J. (2008). 20 Newsgroups. Retrieved November 2, 2018, from http://qwone.com/~jason/20Newsgroups/

Rosario, B., & Hearst, M. a. (2004). Classifying semantic relations in bioscience texts. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 430. https://doi.org/10.3115/1218955.1219010

Rosenberg, A., & Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. *Proceedings of the Joint Conference on*

Universiti Utara Malaysia

Empirical Methods in Natural Language Processing and Computational Natural Language (EMNLP-CoNLL'07), *1*(June), 410–420. https://doi.org/10.7916/D80V8N84

Rumshisky, A. (2008). Resolving Polysemy in Verbs: Contextualized Distributional Approach to Argument Semantics. *Distributional Models of the Lexicon in Linguistics and Cognitive Science, Special Issue of Italian Journal of Linguistics*, 1–27.

Sardar, R. P. ; S. S. ; S. K. N. ; M. M. (2018). Improving Lesk by Incorporating Priority for Word Sense Disambiguation. https://doi.org/10.1109/EAIT.2018.8470436

Sayeed, A., Greenberg, C., & Demberg, V. (2016). Thematic fit evaluation: an aspect of selectional preferences. *ACL 2016, 99.*

Silveira, S. B., & Branco, A. (2012). Combining a double clustering approach with sentence simplification to produce highly informative multi-document summaries. *Proceedings of the 2012 IEEE 13th International Conference on Information Reuse and Integration, IRI 2012*, (1), 482–489. https://doi.org/10.1109/IRI.2012.6303047

Slimani, T. (2013). Description and Evaluation of Semantic Similarity Measures Approaches. *International Journal of Computer Applications*, *80*(10), 25–33. https://doi.org/10.5120/13897-1851

Steinberger, J., & Ježek, K. (2004). Using Latent Semantic Analysis in Text Summarization. *In Proceedings of ISIM 2004*, 93--100.

Sugiyama, M., & Borgwardt, K. (2013). Rapid Distance-Based Outlier Detection via

Sampling. *Advances in Nueral Information Processing Systems 26 (Proceedings of NIPS)*, 1–9.

Sun, F., Guo, J., Lan, Y., Xu, J., & Cheng, X. (2016). Semantic Regularities in Document Representations. Retrieved from http://arxiv.org/abs/1603.07603

Szmeja, P., Ganzha, M., Paprzycki, M., & Pawłowski, W. (2018). Dimensions of Semantic Similarity, 87–125.

Takahashi, T. (2011). Discovering Emerging Topics in Social Streams via Link Anomaly Detection.pdf, *26*, 1–18. https://doi.org/10.1109/icdm.2011.53

Tan, L., Zhang, H., Clarke, C. L. a, & Smucker, M. D. (2015). Lexical Comparison Between Wikipedia and Twitter Corpora by Using Word Embeddings. *Acl*, 657–661.

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). Introduction to Data Mining. *Introduction to Data Mining*, 769.

Torres, S., & Gelbukh, A. (2009). Comparing Similarity Measures for Original WSD Lesk Algorithm. *Advances in Computer Science and Applications*, *43*, 155–166.

Tsai, F. S. (2007). Novelty detection for text documents using named entity recognition. *2007 6th International Conference on Information, Communications & Signal Processing*, (3), 1–5. https://doi.org/10.1109/ICICS.2007.4449883

Turney, P. D., & Pantel, P. (2010). ★★★★★From Frequency to Meaning_ Vector Space Models of Semantics（讲的非常好，但是我还只看了三分之一）.pdf, *37*,

141–188. https://doi.org/10.1613/jair.2934

Upadhyaya, S., & Singh, K. (2012). Classification based outlier detection techniques. *Int J Comput Trends Technol*, *3*, 294–298. Retrieved from http://www.ijcttjournal.org/Volume3/issue-2/IJCTT-V3I2P118.pdf

Wagner, A. (2000). Enriching a lexical semantic net with selectional preferences by means of statistical corpus analysis. *Proceedings of ECAI Workshop on Ontology Learning and Population*, 37–42. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Enriching+a+Lexical+Semantic+Net+with+Selectional+Preferences+by+Means+of+Statistical+Corpus+Analysis#0

Wang, Y., Ni, X., Sun, J.-T., Tong, Y., & Chen, Z. (2011). Representing document as dependency graph for document clustering. *Proceedings of the 20th ACM International Conference on Information and Knowledge Management - CIKM '11*, 2177. https://doi.org/10.1145/2063576.2063920

Wehmeier, K. F. (2004). Wittgensteinian Predicate Logic. *Notre Dame Journal of Formal Logic*, *45*(1), 1–11. https://doi.org/10.1305/ndjfl/1094155275

William Wei Song, Chenlu Lin, A. F. (2017). An Euclidean similarity measurement approach for hotel rating data analysis. Retrieved from https://ieeexplore.ieee.org/abstract/document/7951927/authors

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). A biterm topic model for short texts. *WWW '13 Proceedings of the 22nd International Conference on World Wide Web*,

Universiti Utara Malaysia

1445–1456. Retrieved from http://dl.acm.org/citation.cfm?id=2488388.2488514

Yang, Y., Zhang, J., Carbonell, J., & Jin, C. (2002). Topic-conditioned novelty detection. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '02*, 688. https://doi.org/10.1145/775047.775150

Yih, W., & Meek, C. (n.d.). Improving Similarity Measures for Short Segments of Text, 1489–1494.

Yin, J., & Wang, J. (2016). A Model-based Approach for Text Clustering with Outlier Detection. *Icde*, 625–636. https://doi.org/10.1109/ICDE.2016.7498276

Yoo, J., & Yang, D. (2015). Classification Scheme of Unstructured Text Document using TF-IDF and Naive Bayes Classifier Text Classification using TF-IDF and Naï ve Bayes Classifier, *111*(Comcoms), 263–266. https://doi.org/10.14257/astl.2015.111.50

Yuhanis, S. S. kamaruddin and Y. (2015). constructing canonical data model for text document clustering, 4.

Zhang, D., Zhai, C., Han, J., Srivastava, A., & Oza, N. (2009). Topic modeling for OLAP on multidimensional text databases: Topic cube and its applications. *Statistical Analysis and Data Mining*, *2*(5–6), 378–395. https://doi.org/10.1002/sam.10059

Zhang, W., Tang, X., & Yoshida, T. (2015). TESC: An approach to TExt classification using Semi-supervised Clustering. *Knowledge-Based Systems*, *75*, 152–160.

Universiti Utara Malaysia

https://doi.org/10.1016/j.knosys.2014.11.028

Zhang, W., Xiao, F., Li, B., & Zhang, S. (2016). Using SVD on Clusters to Improve Precision of Interdocument Similarity Measure. *Computational Intelligence and Neuroscience*, *2016*. https://doi.org/10.1155/2016/1096271

Zhang, Z. Z. Z., & Feng, X. F. X. (2009). New Methods for Deviation-Based Outlier Detection in Large Database. *2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery*, *1*. https://doi.org/10.1109/FSKD.2009.303

Zhou, G., Zhao, J., Liu, K., & Cai, L. (2011). Exploiting Web-Derived Selectional Preference to Improve Statistical Dependency Parsing. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 1556–1565.

Zhou, Y., Fleischmann, K. R., & Wallace, W. A. (2010). Automatic text analysis of values in the enron email dataset: Clustering a social network using the value patterns of actors. *Proceedings of the Annual Hawaii International Conference on System Sciences*, 1–10. https://doi.org/10.1109/HICSS.2010.77

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, *39*(4), 561–577. https://doi.org/ROC; Receiver-Operating Characteristic; SDT; Signal Detection Theory

Universiti Utara Malaysia

LIST OF PUBLICATIONS WITH RESEARCH & GRANT WORK

2015 A Framework For Semantic-based Anomaly Detection In Text, 4th International Conference on Internet Applications, Protocols and Services*(NETAPP)*, Malaysia December 1-3, 2015.
2015 Expert Directory System for Managing Organizational Knowledge
2016 Representing Semantics of Text By Acquiring Its Canonical Form, 3rd International Multi-conference on Artificial Intelligence Technology *(M-CAIT 2016)*, Bangi, Selangor Malaysia August 23-24, 2016.
2017 Representing Semantics of Text by Acquiring its Canonical Form
2017 Framework for Enhancing A Wearable Device that Converts Sound, Text and Image Into Automatic Sign Language Recognizing System (ASLR)
2017 Framework on comparative analysis of Text Representation Schemes and Similarity Measures For Sentences
2017 Combined Word Sense Disambiguation Algorithms with Latent Semantic Analysis to identify semantic similarity in unstructured textual data
2017 Visualization of Spoken Language for Deaf People
2018 Graph-based Representation for Sentence Similarity Measure: A Comparative Analysis

List of papers in-view

- Extraction of Agro-food terms from online news website in Malaysia
- Integration of Word sense disambiguation algorithms to analyze and identify similar terms in documents
- Optimizing sequential exception techniques for anomaly detection in corpuse
- A systematic review on text anomaly (from all levels of text; word, sentence, document, events and topics)

Universiti Utara Malaysia

APPENDIX A

Process Flow in ESET

As described in section 3.3 This study centres on the enhancement of ESET following this flow.

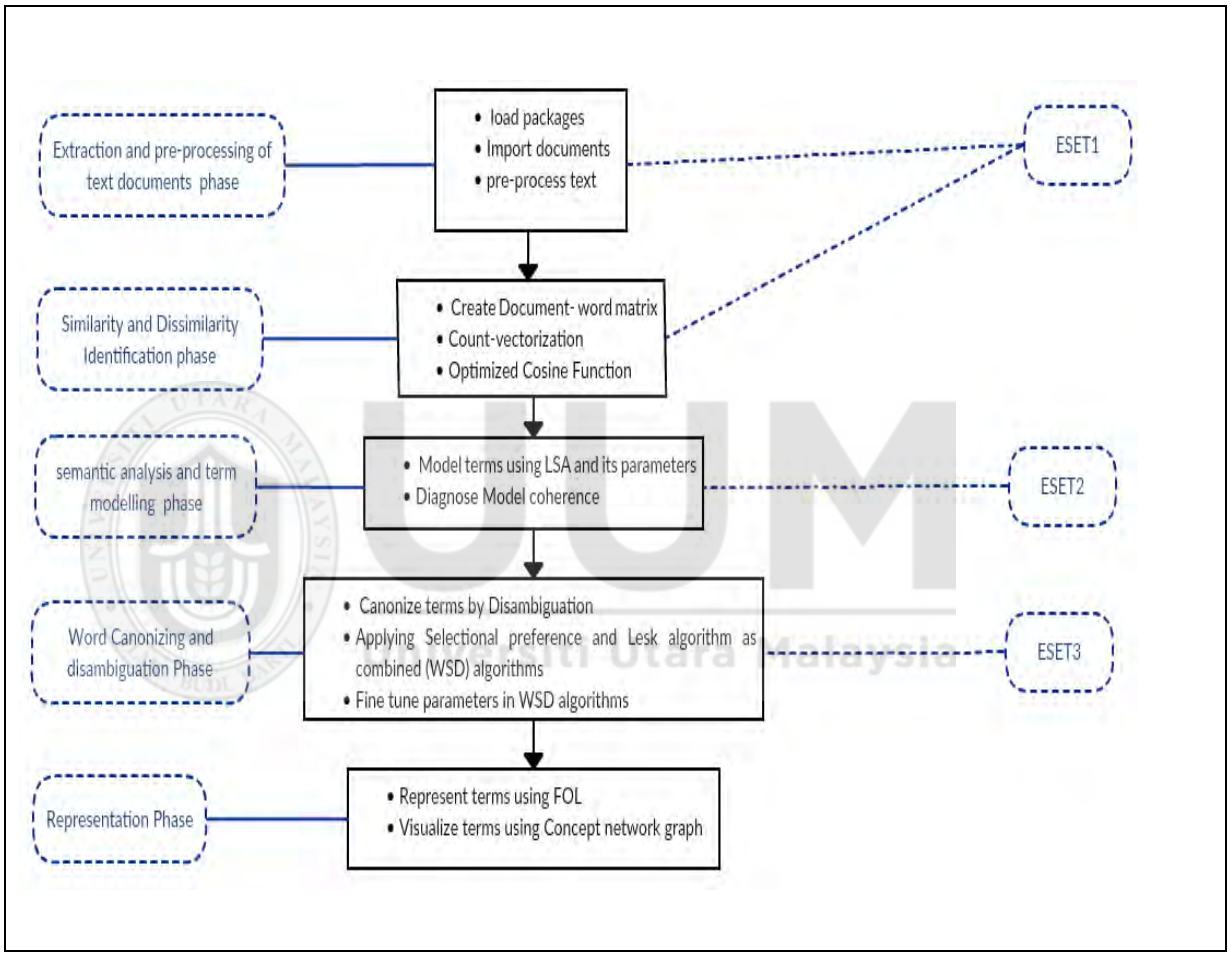

Figure A1. ESET Flowchart

APPENDIX B

Code Snippet of Results Extracted from ENRON POI

As described in section 4.2 in chapter four an experiment was carried out to assess the feasibility of the ESET. In this experiment, sentences were extracted and tested from Mail messages sent and received from Kenneth Lay. Figure A1 shows why and how Kenneth Lay was chosen as a POI.

Data Preparation Phase

```
# ### Importing libraries and magics
# ### Import the file which contain the data to our variable
# In[3]:
# Load the dictionary containing the data
with open(os.getcwd()+"/final_project_data.pkl", "rb") as data_file:
    data_init = pickle.load(data_file)
# ### Converting the data from a python dictionary to a pandas dataframe
# In[4]:
#Converting the data from a python dictionary to a pandas dataframe
data_df = pd.DataFrame.from_dict(data_init, orient='index')
raw_data = data_df.copy()
# #### Now check the structure of the new data frame to find out how
many total number of observation and column are prsent
# In[5]:
data_df.count().sort_values()
# In[9]:
#dropping 'poi' and 'email_address' variables
data_df = data_df.drop(["email_address"], axis=1)
data_temp = data_df.drop(["poi"], axis=1)
data_temp[data_temp.isnull().all(axis=1)]
        ys = dataframe[[feature]]
        quartile_1, quartile_3 = np.percentile(ys, [25, 75])
        iqr = quartile_3 - quartile_1
        lower_bound = int(round(quartile_1 - (iqr * 3)))
        upper_bound = int(round(quartile_3 + (iqr * 3)))
        partial_result = list(np.where((ys > upper_bound) | (ys <
lower_bound))[0])
        print(feature, len(partial_result))
        result.update(partial_result)
    print("Total number of records with extreme values: " +
    selector = SelectPercentile(percentile=100)
    a = selector.fit(X, y)
    plt.figure(figsize=(12,9))
    sns.barplot(y=X.columns, x=a.scores_)
# In[40]:
plot_importance(data)
```

Universiti Utara Malaysia

SET Phase

```
# coding: utf-8
# ### Sequential Exception Technique (SET)
# Identify the POIs using SET and print their names.
SET_data.head()
# In[46]:
cols = [ 'salary', 'bonus', 'long_term_incentive', 'deferred_income',
def SET(m,SET_data):
# Set the value of parameter m = the no. of iterations you require
    Card = pd.Series(np.NAN)
    DS=pd.Series(np.NAN)
    idx_added = pd.Series(np.NAN)
    pos = 0
    for j in range(1,m+1):
        new_indices                                                    =
np.random.choice(e_names.index,len(e_names),replace=False)
        for i in pd.Series(new_indices).index:
            idx_added[i+pos] = new_indices[i]
 DS[i+pos]=sum(np.var(SET_data.loc[e_names[new_indices[:i+1]]]))
            Card[i+pos] = len(e_names[:i+1])
        pos = pos+i+1
    df = pd.DataFrame({'Index_added':idx_added,'DS':DS,'Card':Card})
    df ['DS_Prev'] = df.DS.shift(1)
    df['Card_prev'] = df.Card.shift(1)
    df.Card_prev[(df.Card == 1)] = 0
    df = df.fillna(0)
    df['Smoothing'] = (df.Card - df.Card_prev)*(df.DS - df.DS_Prev)
    # find indexes of sets with max sf
    maxsf = []
    for i in range(len(df.DS)):
        if df.Smoothing[i] == df.Smoothing.max():
            maxsf.append(i)
    #print(maxsf)
    N = len(e_names)
    excp_set = []
    for i in range(len(maxsf)):
        j = maxsf[i]
        k=j+1
        temp = []
        temp.append(df.Index_added[j])
        excp_set.append(temp.copy())
        temp_prev = pd.DataFrame()
        temp_j = pd.DataFrame()
        a=j
        while(a%N!=0):
            temp_row = SET_data.loc[e_names[df.Index_added[a]]]
            temp_j = temp_j.append(temp_row)
            a=a-1
        temp_row = SET_data.loc[e_names[df.Index_added[a]]]
        temp_j = temp_j.append(temp_row)
        temp_prev = temp_j.copy()                # Ij-1
```

Universiti Utara Malaysia

```
        temp_prev.drop(temp_prev.index[0],inplace=True)
        #temp_prev.index = np.arange(len(temp_prev))
        while(k%N!=0):
            K_element = SET_data.loc[e_names[df.Index_added[k]]]   # K th
element
            temp_prev = temp_prev.append(K_element)          # Ij-1 U {ik}
            temp_j = temp_j.append(K_element)             # Ij U {ik}
            Dk0 = sum(np.var(temp_prev)) - df.DS[j-1]
            Dk1 = sum(np.var(temp_j)) - df.DS[j]
            if Dk0-Dk1 >= df.DS[j]:                    # If Dk0 - Dk1 >= Dj
excp_set[i].append(df.Index_added[k])
            temp_prev.drop(temp_prev.index[len(temp_prev)-
1],inplace=True)
            temp_j.drop(temp_j.index[len(temp_j)-1],inplace=True)
            k+=1
    #print(excp_set)                          # contains the indices of exception
elements.
    return excp_set
# In[ ]:
excp_set = SET(1000,SET_data)
# In[ ]:
# Printing the POIs.
print("\nException set: \n")
for i in range(len(excp_set)):
    print(e_names[excp_set[i]])
```

Figure A1 code snippet of POI with the most Total Payment Information using SET.

Universiti Utara Malaysia

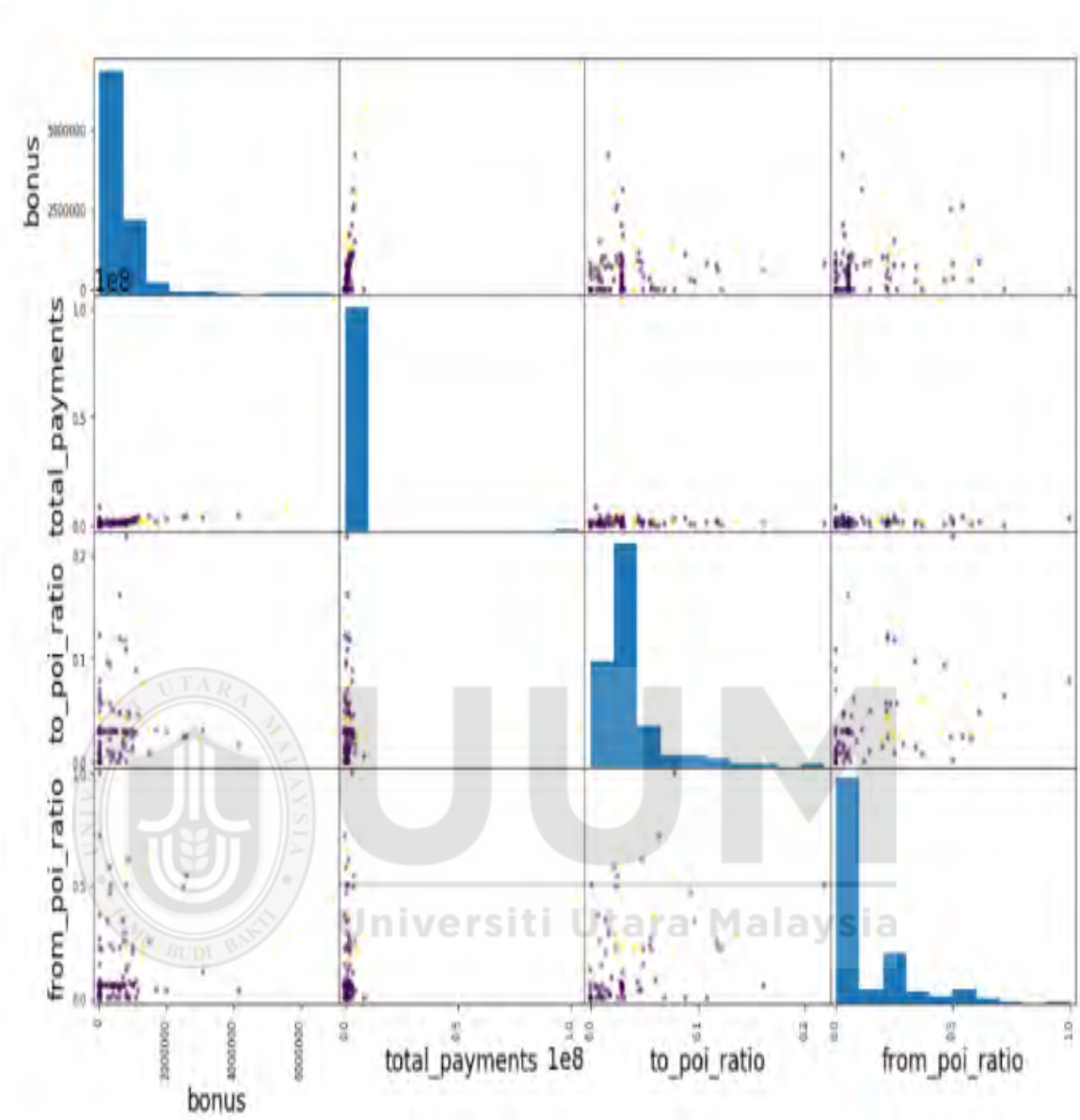

Figure B1 Data pre-processing Phase

Figure A2 shows a scatter matrix with an overall visualization of the ENRON email financial data.

OUTPUT	
FREVERT MARK A	12
BELDEN TIMOTHY N	9
SKILLING JEFFREY K	9
BAXTER JOHN C	8
LAVORATO JOHN J	8
DELAINEY DAVID W	7
KEAN STEVEN J	7
HAEDICKE MARK E	7
WHALLEY LAWRENCE G	7
RICE KENNETH D	6
KITCHEN LOUISE	6
LAY KENNETH L	**15**

Figure A3 presents the Output of SET codes.

After identifying the most POI (Kenneth Lay) by comparing Total payment information as seen in the figures shown above. To identify other POI who mail and received messages from Kenneth Lay, some pseudocodes were also developed to identify other POIs. In addition, pseudocodes used in extracted mail messages from the identified POI.

Analysing ENRON and Identifying most sent and received Mails

```
import os
from collections import Counter
from email.parser import Parser
rootdir = "C:\\Users\\Shantnu\\Desktop\\Data Sources\\maildir\\"
def email_analyse(inputfile, to_email_list, from_email_list,
email_body):
    with open(inputfile, "r") as f:
        data = f.read()
    email = Parser().parsestr(data)
    if email['to']:
        email_to = email['to']
        email_to = email_to.replace("\n", "")
        email_to = email_to.replace("\t", "")
        email_to = email_to.replace(" ", "")
        email_to = email_to.split(",")
        for email_to_1 in email_to:
            to_email_list.append(email_to_1)
    from_email_list.append(email['from'])
to_email_list = []
from_email_list = []
email_body = []
for directory, subdirectory, filenames in  os.walk(rootdir):
    for filename in filenames:
```

Word Frequency

```
        email_analyse(os.path.join(directory, filename), to_email_list,
from_email_list, email_body )
print("\nTo email adresses: \n")
print(Counter(to_email_list).most_common(10))
print("\nFrom email adresses: \n")
print(Counter(from_email_list).most_common(10))
import os
from collections import Counter
from email.parser import Parser
rootdir = "C:\\Users\\Shantnu\\Desktop\\Data Sources\\maildir\\"
def email_analyse(inputfile, to_email_list, from_email_list,
email_body):
    with open(inputfile, "r") as f:
        data = f.read()
    email = Parser().parsestr(data)
    if email['to']:
        email_to = email['to']
        email_to = email_to.replace("\n", "")
        email_to = email_to.replace("\t", "")
        email_to = email_to.replace(" ", "")
        email_to = email_to.split(",")
        for email_to_1 in email_to:
            to_email_list.append(email_to_1)
    from_email_list.append(email['from'])
to_email_list = []
from_email_list = []
email_body = []
for directory, subdirectory, filenames in  os.walk(rootdir):
    for filename in filenames:
        email_analyse(os.path.join(directory, filename), to_email_list,
from_email_list, email_body )
print("\nTo email adresses: \n")
print(Counter(to_email_list).most_common(10))
print("\nFrom email adresses: \n")
print(Counter(from_email_list).most_common(10))
```

Figure A4 presents a code snippet for extracting and analysing mail messages sent and received from POIs

Mail messages of POIs were all analysed and extracted. These mail messages contain other attributes like senders and recipients ID dates and the mime version. In this study, we are only interested in extracting the body of mail messages to avoid unnecessary content and as well reduce the workload of text data pre-processing.

Universiti Utara Malaysia

APPENDIX C

A Sample of Most Frequent Terms Using ESET

No	Term1	Term2	Term 3	Term 4
1	Enron	Time	Please	Deal
2	Business	Thank	Thank	Gas
3	manage	Day	Attach	Price
4	Meet	Don't	Email	Contract
5	Market	Call	Enron	Power
6	Company	Talk	Call	Rate
7	Vince	Hope	Copying	Trade
8	Report	Ill	Fax	Day
9	Time	Bit	File	Month
10	Energy	Trying	Message	Companies
11	Information	Guy	Information	Energy
12	Please	Night	Phone	Transaction
13	Trade	Friday	Send	Product
14	Discuss	Weekend	Corp	Term
15	Regards	Love	Kay	Custom
16	Team	Item	Receive	Cost
17	Plan	Email	Question	Thank
18	Service	people	Draft	Purchase
19	Message	File	Price	Organization
20	Phone	Information	Business	Electricity

Figure A5 presents list of ENRON Terms

Universiti Utara Malaysia

Figure A5 is a list of terms from ENRON using the ESET2 of the study research design. Daily Kos Bloggs. The study also presents list of some terms from the Daily Kos blogs data using the word-cloud as a visualization scheme